



A lorikeet is a small to medium-sized parrot with a brightly colored plumage.

Prompt Generation for Zero-Shot Image Classification

Sarah Pratt

sarahpratt.github.io

Classify this dog!



Classify this dog!



A photo of a saluki



A photo of a vizsla



A photo of a Ibizan hound

Classify this dog!



The easiest way to identify a Saluki is by its iconic long, silky ears.



A Vizsla is a short-haired, red-brown hunting dog.



The Ibizan Hound is a slender, elegant dog with large, bat-like ears.

Classify this dog!



The easiest way to identify a Saluki is by its iconic long, silky ears.



A Vizsla is a short-haired, red-brown hunting dog.



The Ibizan Hound is a slender, elegant dog with large, bat-like ears.

Classify this dog!



The easiest way to identify a Saluki is by its iconic long, silky ears.



LLM Generated



A Vizsla is a short-haired, red-brown hunting dog.

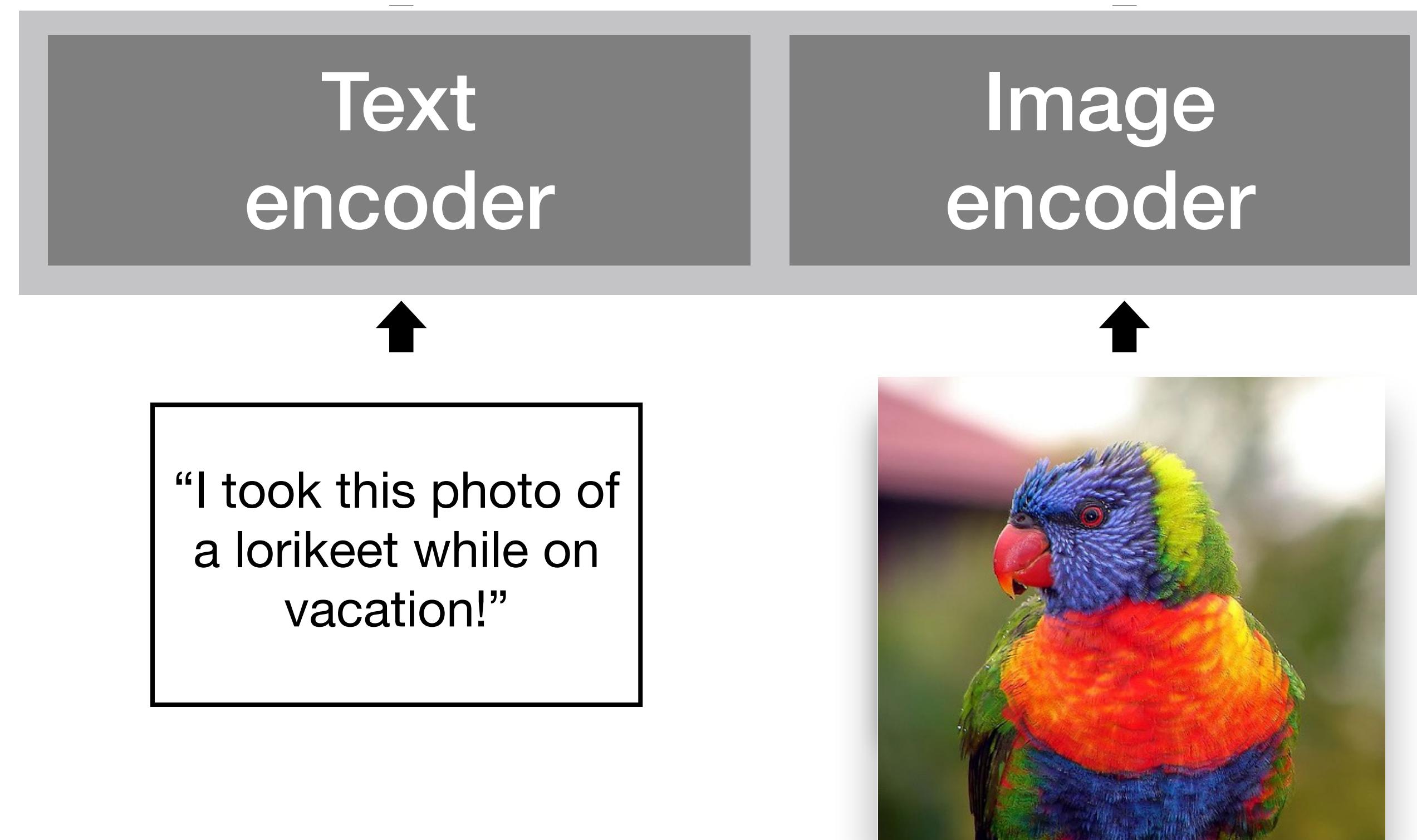


The Ibizan Hound is a slender, elegant dog with large, bat-like ears.

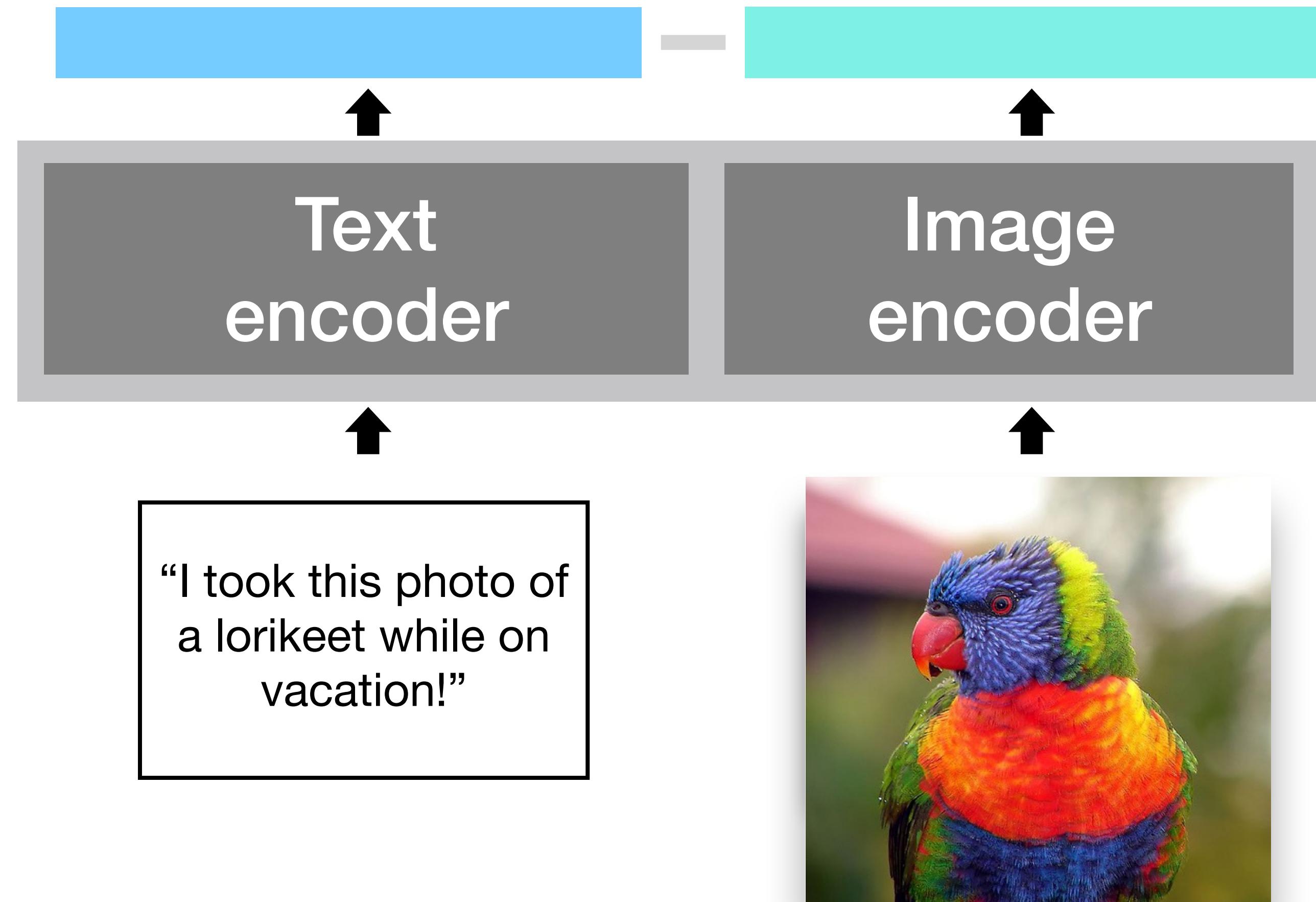
CLIP Training

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

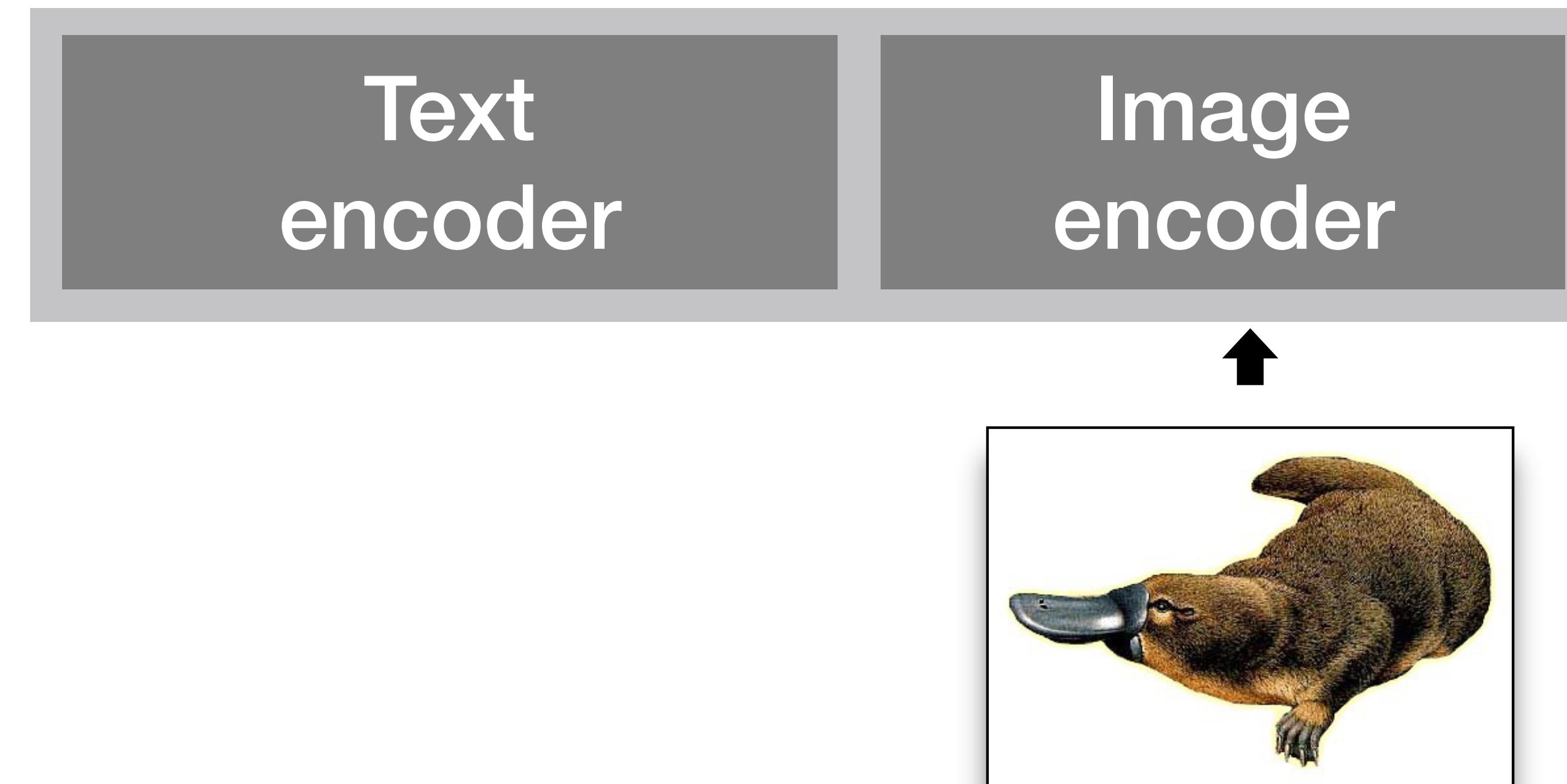
CLIP Training



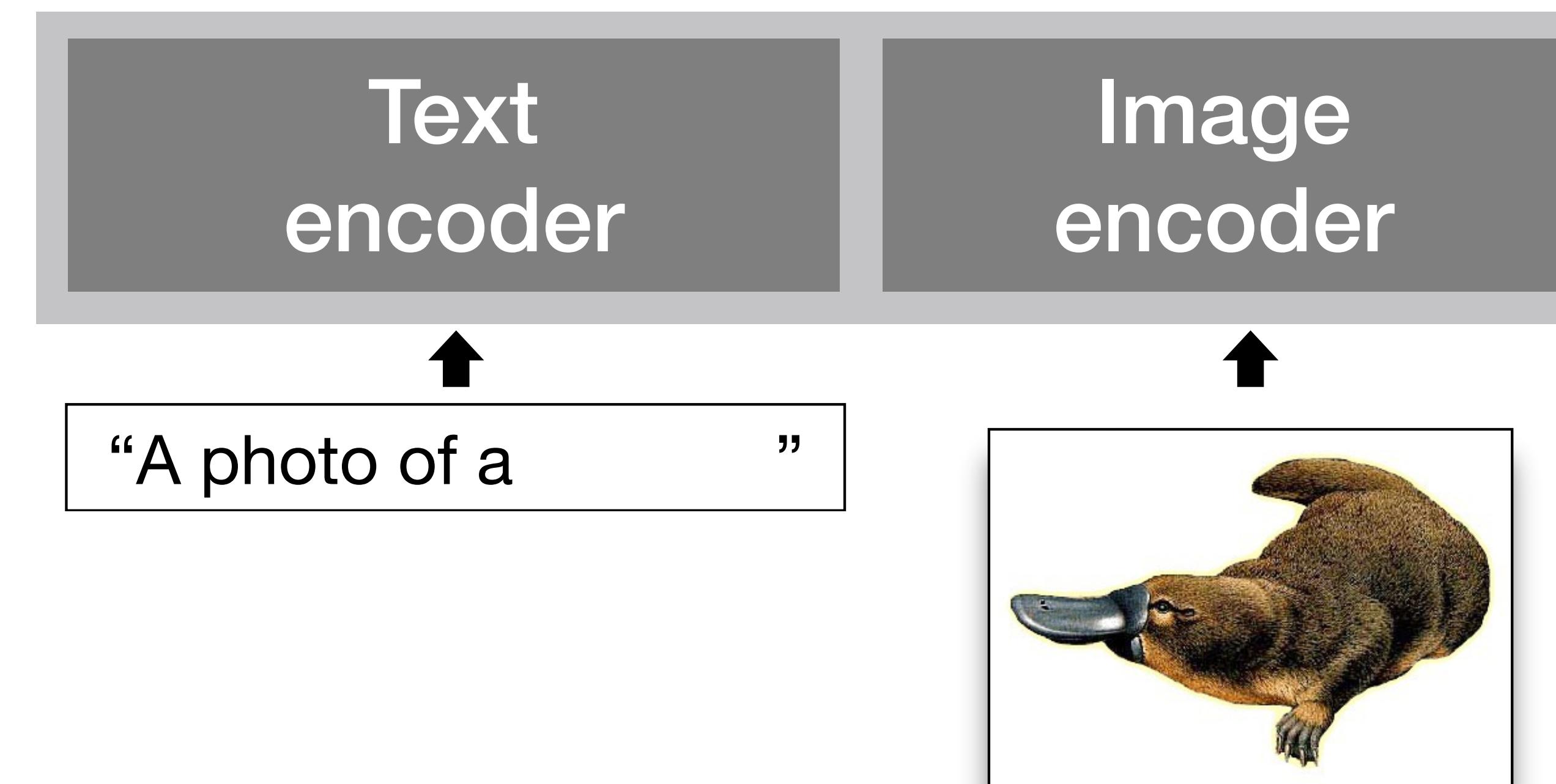
CLIP Training



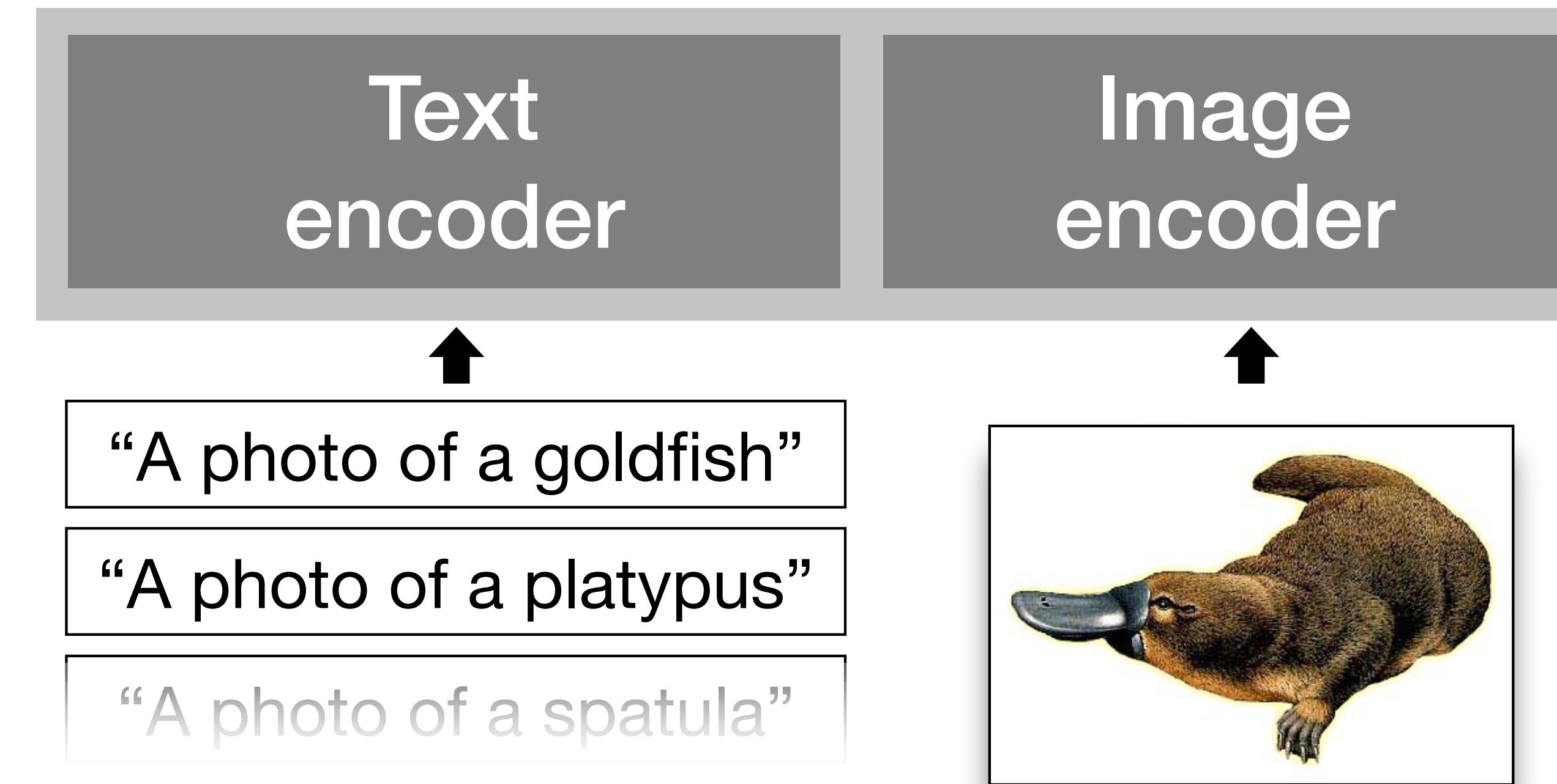
CLIP Inference



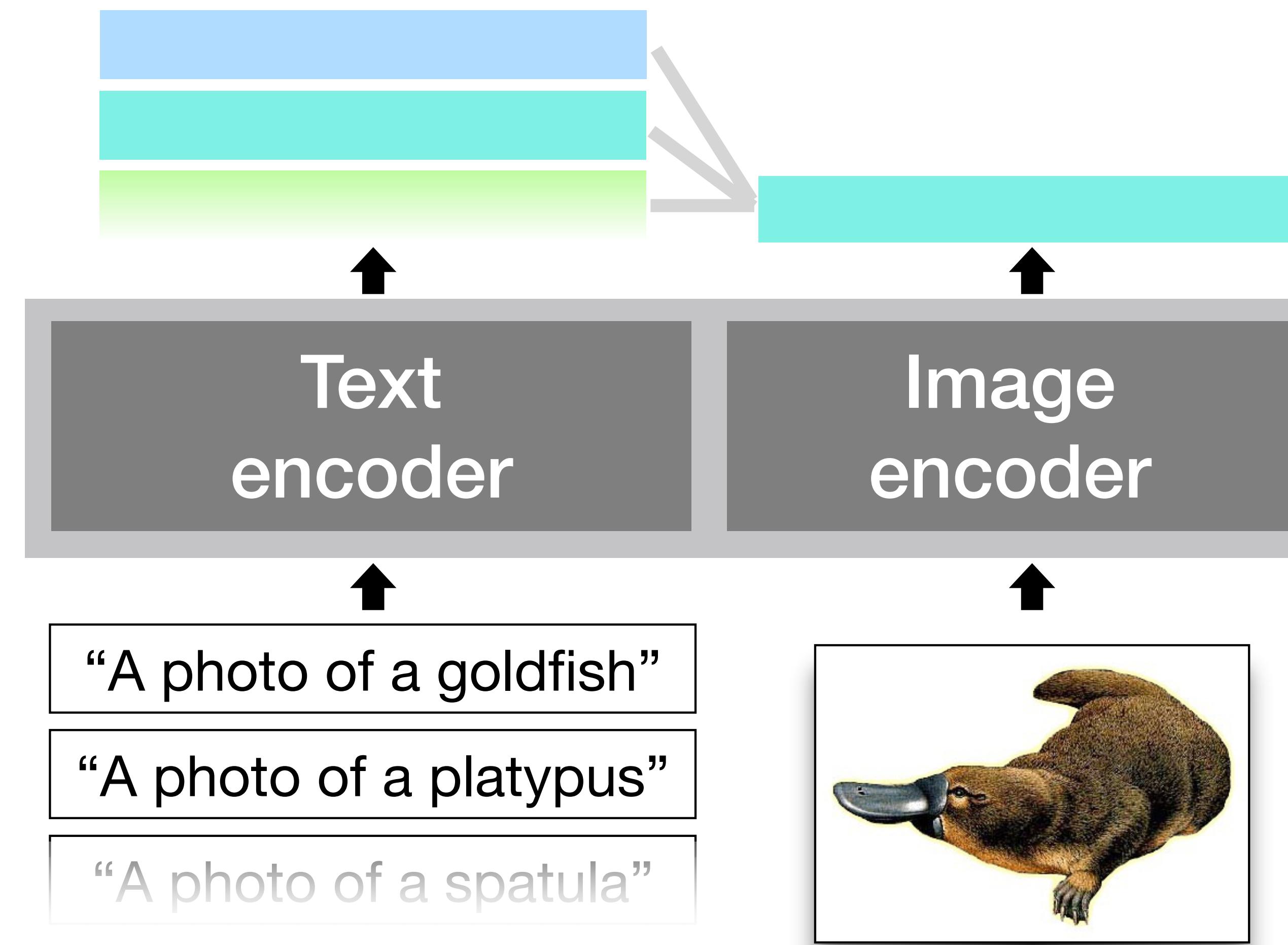
CLIP Inference



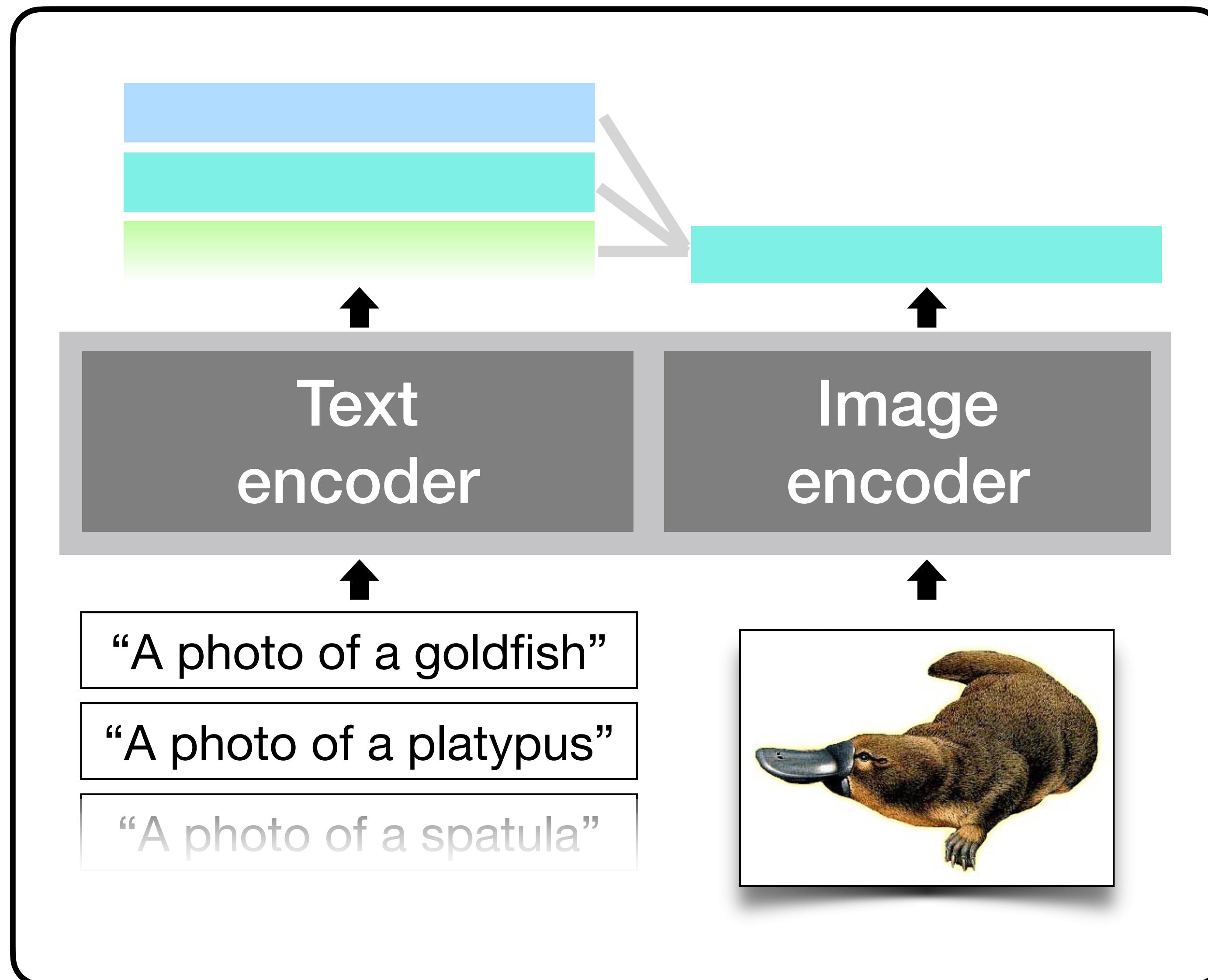
CLIP Inference



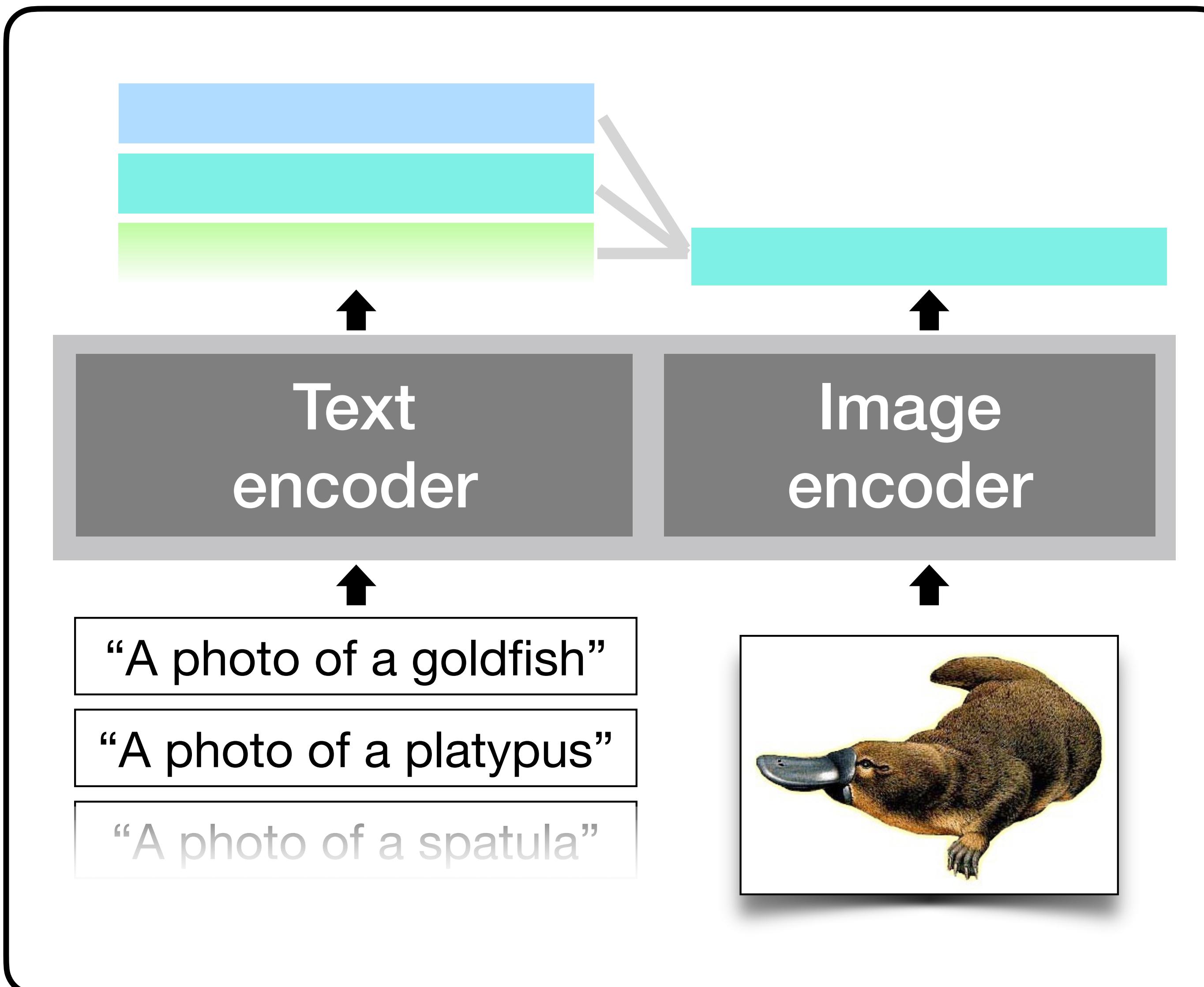
CLIP Inference



Standard Zero-shot

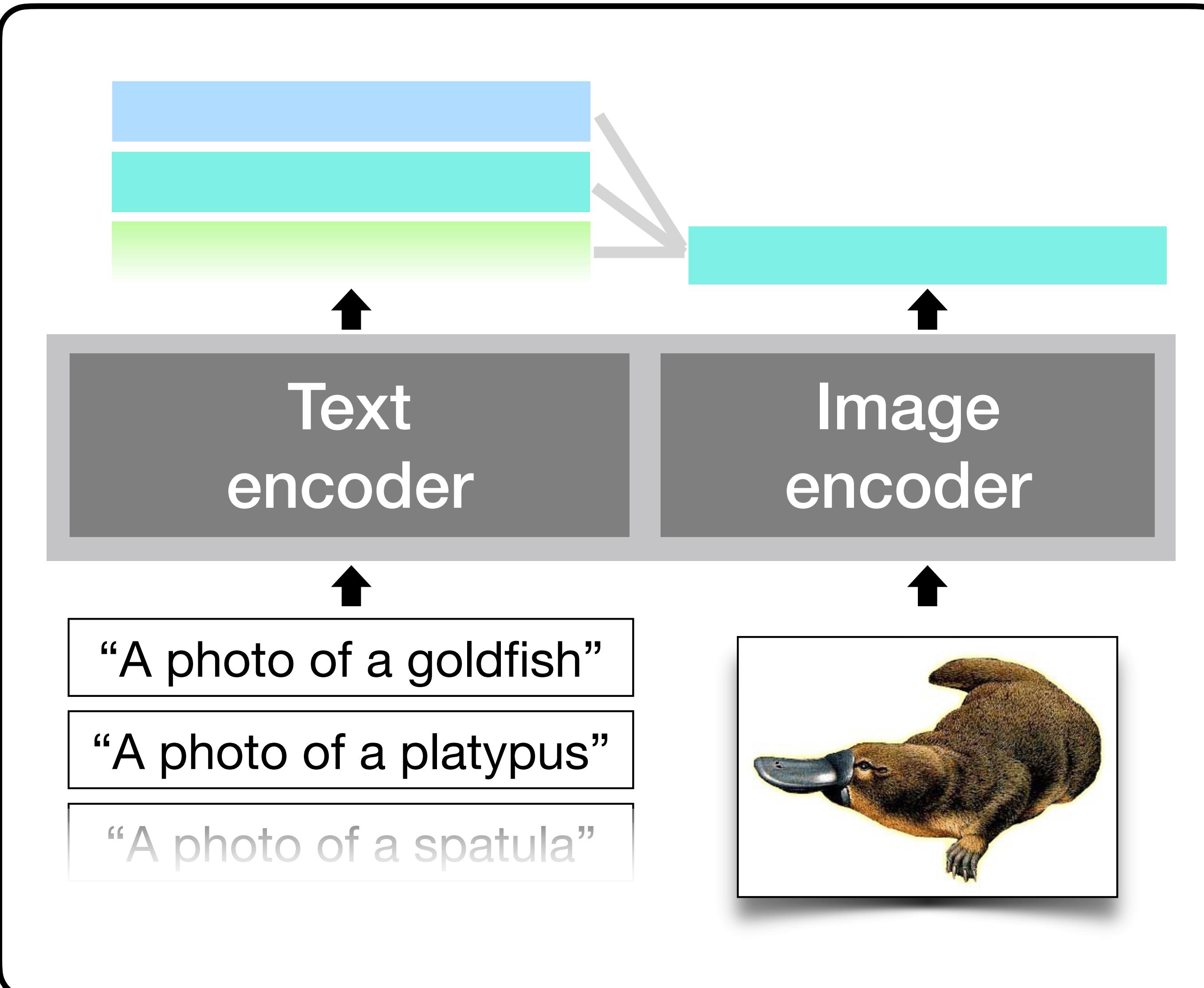


Standard Zero-shot



Drawbacks

Standard Zero-shot



Drawbacks

1. No specific visual information

1. No specific visual information



The easiest way to identify a Saluki is by its iconic long, silky ears.



A Vizsla is a short-haired, red-brown hunting dog.



The Ibizan Hound is a slender, elegant dog with large, bat-like ears.

1. No specific visual information



A photo of a saluki



A photo of a vizsla



A photo of a ibizan hound

1. No specific visual information



A photo of a saluki

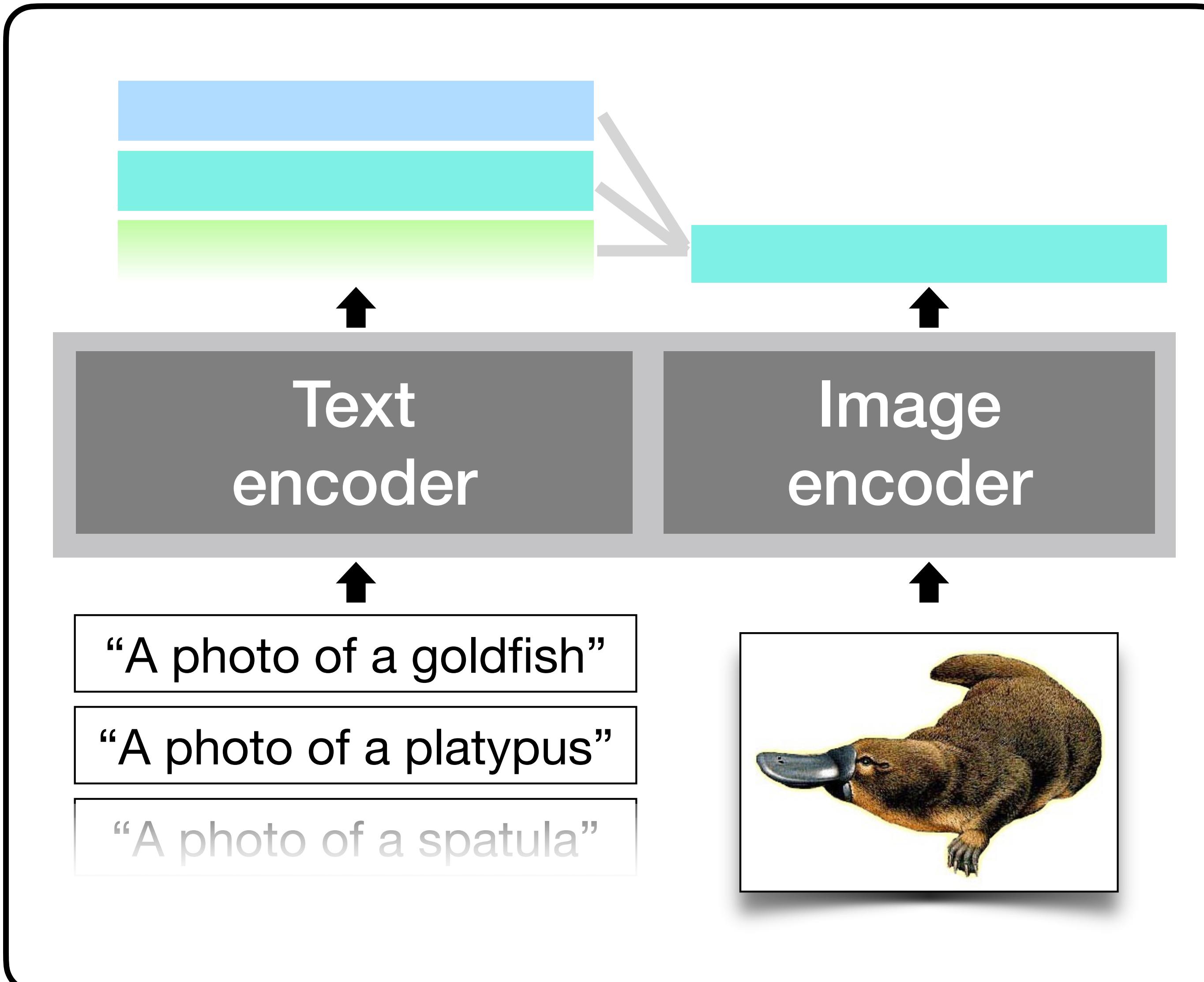


A photo of a vizsla



A photo of a ibizan hound

Standard Zero-shot



Drawbacks

- 1. No specific visual information**
- 2. Many hand-written prompts**

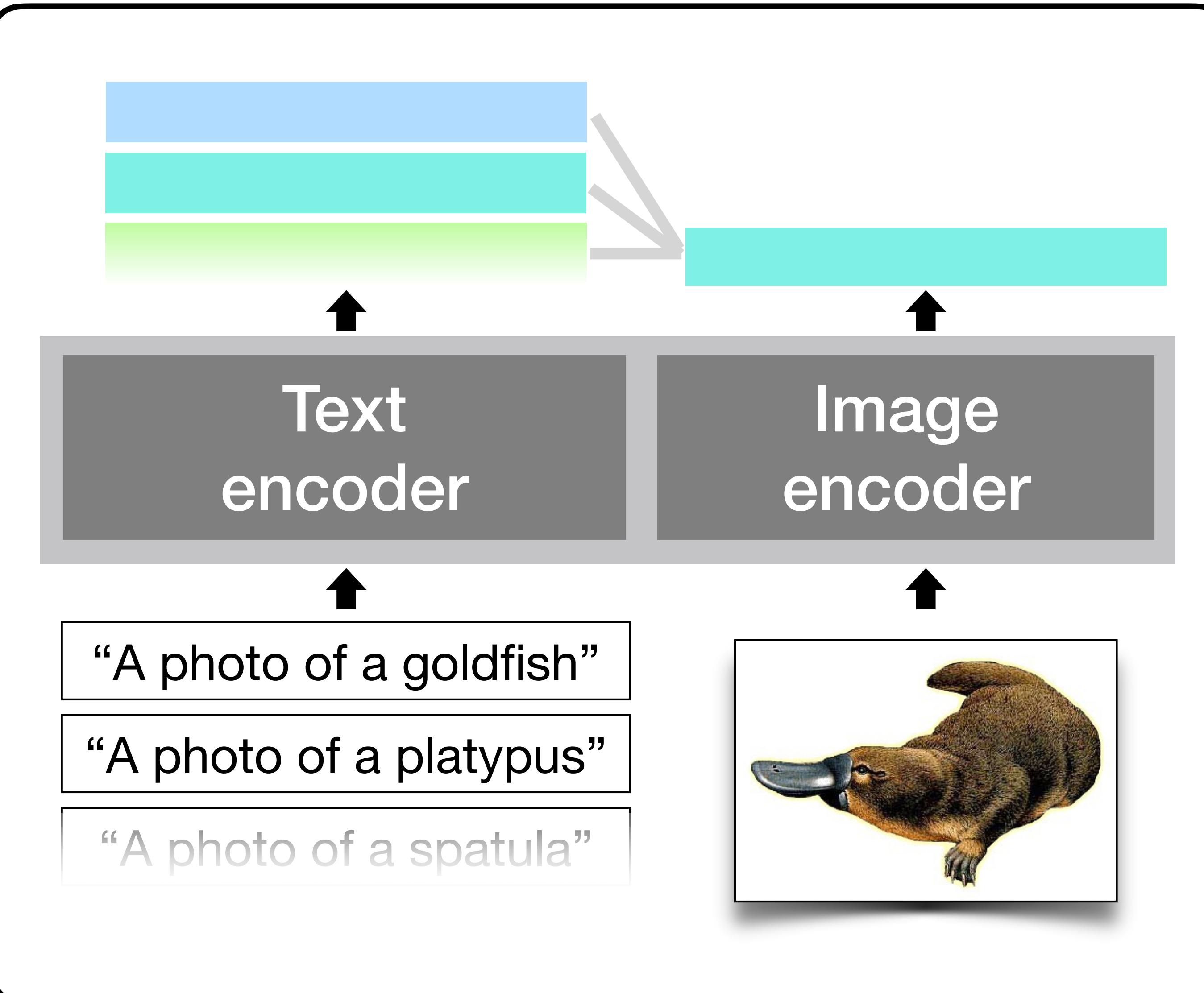
2. Many hand-written prompts

a bad photo of a {}.
a photo of many {}.
a sculpture of a {}.
a photo of the hard to see {}.
a low resolution photo of the {}.
a rendering of a {}.
graffiti of a {}.
a bad photo of the {}.
a cropped photo of the {}.
a tattoo of a {}.
the embroidered {}.
a photo of a hard to see {}.
a bright photo of a {}.
a photo of a clean {}.
a photo of a dirty {}.
a dark photo of the {}.
a drawing of a {}.
a photo of my {}.
the plastic {}.
a photo of the cool {}.
a close-up photo of a {}.
a black and white photo of the {}.
a painting of the {}.
a painting of a {}.

a pixelated photo of the {}.
a sculpture of the {}.
a bright photo of the {}.
a cropped photo of a {}.
a plastic {}.
a photo of the dirty {}.
a jpeg corrupted photo of a {}.
a blurry photo of the {}.
a photo of the {}.
a good photo of the {}.
a rendering of the {}.
a {} in a video game.
a photo of one {}.
a doodle of a {}.
a close-up photo of the {}.
a photo of a {}.
the origami {}.
the {} in a video game.
a sketch of a {}.
a doodle of the {}.
a origami {}.
a low resolution photo of a {}.
the toy {}.
a rendition of the {}.

a photo of the clean {}.
a photo of a large {}.
a rendition of a {}.
a photo of a nice {}.
a photo of a weird {}.
a blurry photo of a {}.
a cartoon {}.
art of a {}.
a sketch of the {}.
a embroidered {}.
a pixelated photo of a {}.
itap of the {}.
a jpeg corrupted photo of the {}.
a good photo of a {}.
a plushie {}.
a photo of the nice {}.
a photo of the small {}.
a photo of the weird {}.
the cartoon {}.
art of the {}.
a drawing of the {}.
a photo of the large {}.
a black and white photo of a {}.
the plushie {}.

Standard Zero-shot



Drawbacks

1. No specific visual information
2. Many hand-written prompts
3. Contain information about data distribution

3. Contain information about data distribution

- | | | |
|------------------------------------|---------------------------------|-----------------------------------|
| a bad photo of a {}. | a pixelated photo of the {}. | a photo of the clean {}. |
| a photo of many {}. | a sculpture of the {}. | a photo of a large {}. |
| a sculpture of a {}. | a bright photo of the {}. | a rendition of a {}. |
| a photo of the hard to see {}. | a cropped photo of a {}. | a photo of a nice {}. |
| a low resolution photo of the {}. | a plastic {}. | a photo of a weird {}. |
| a rendering of a {}. | a photo of the dirty {}. | a blurry photo of a {}. |
| graffiti of a {}. | a jpeg corrupted photo of a {}. | a cartoon {}. |
| a bad photo of the {}. | a blurry photo of the {}. | art of a {}. |
| a cropped photo of the {}. | a photo of the {}. | a sketch of the {}. |
| a tattoo of a {}. | a good photo of the {}. | a embroidered {}. |
| the embroidered {}. | a rendering of the {}. | a pixelated photo of a {}. |
| a photo of a hard to see {}. | a {} in a video game. | itap of the {}. |
| a bright photo of a {}. | a photo of one {}. | a jpeg corrupted photo of the {}. |
| a photo of a clean {}. | a doodle of a {}. | a good photo of a {}. |
| a photo of a dirty {}. | a close-up photo of the {}. | a plushie {}. |
| a dark photo of the {}. | a photo of a {}. | a photo of the nice {}. |
| a drawing of a {}. | the origami {}. | a photo of the small {}. |
| a photo of my {}. | the {} in a video game. | a photo of the weird {}. |
| the plastic {}. | a sketch of a {}. | the cartoon {}. |
| a photo of the cool {}. | a doodle of the {}. | art of the {}. |
| a close-up photo of a {}. | a origami {}. | a drawing of the {}. |
| a black and white photo of the {}. | a low resolution photo of a {}. | a photo of the large {}. |
| a painting of the {}. | the toy {}. | a black and white photo of a {}. |
| a painting of a {}. | a rendition of the {}. | the plushie {}. |

3. Contain information about data distribution

a photo of the {}.

the toy {}.

3. Contain information about data distribution

a photo of the {}.

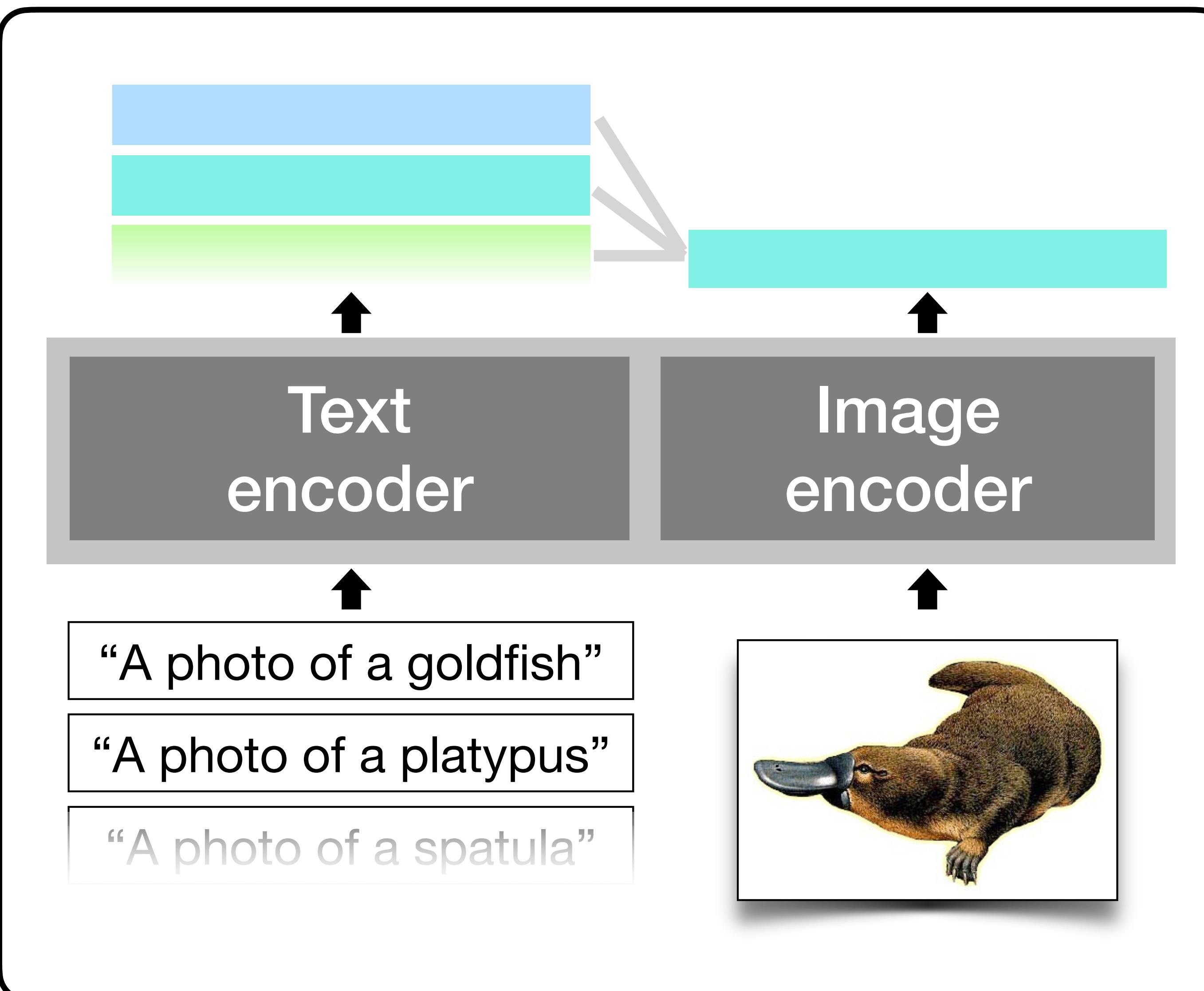
the toy {}.



Pirate Ship

Triceratops

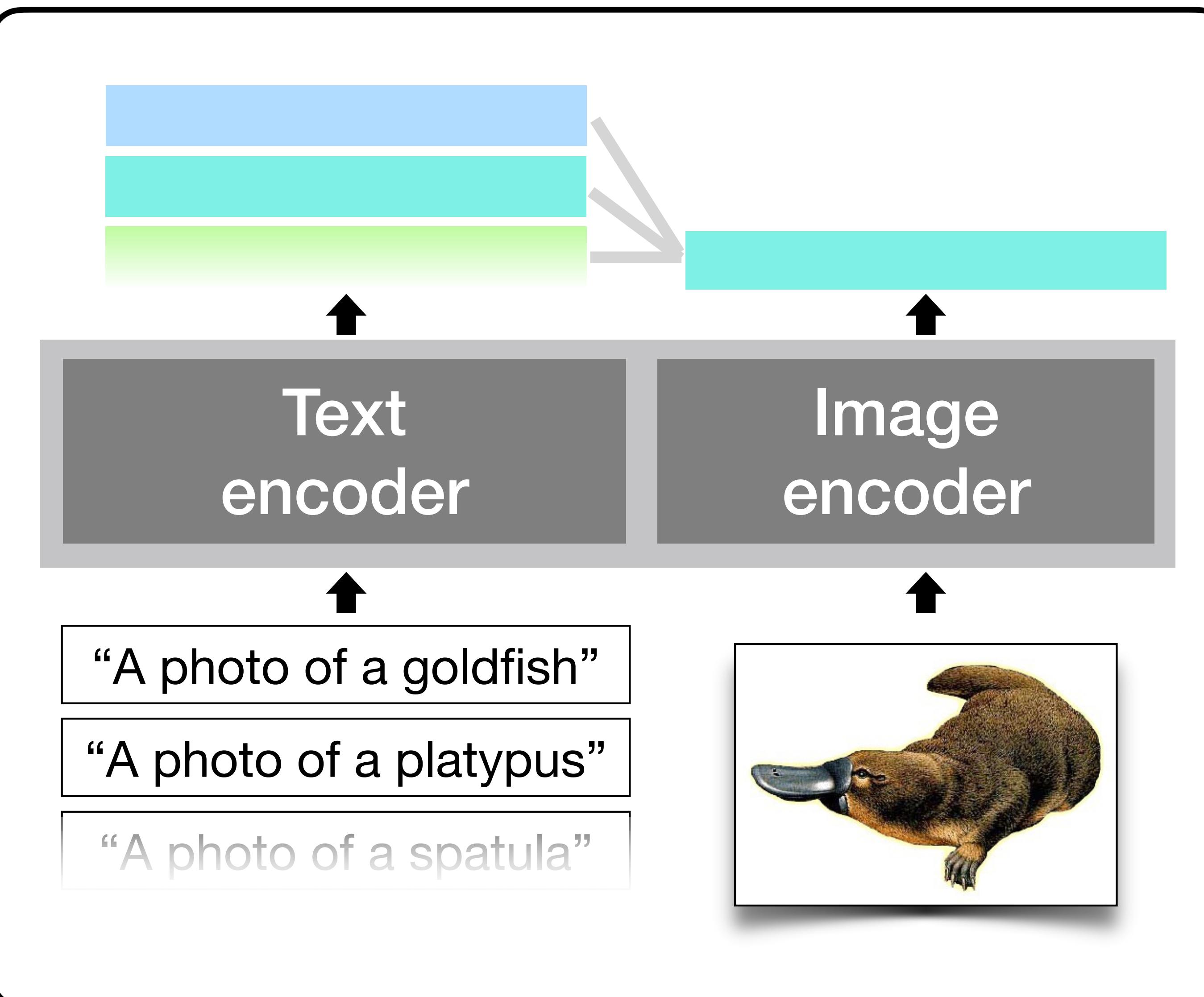
Standard Zero-shot



Drawbacks

1. No specific visual information
2. Many hand-written prompts
3. Contain information about data distribution

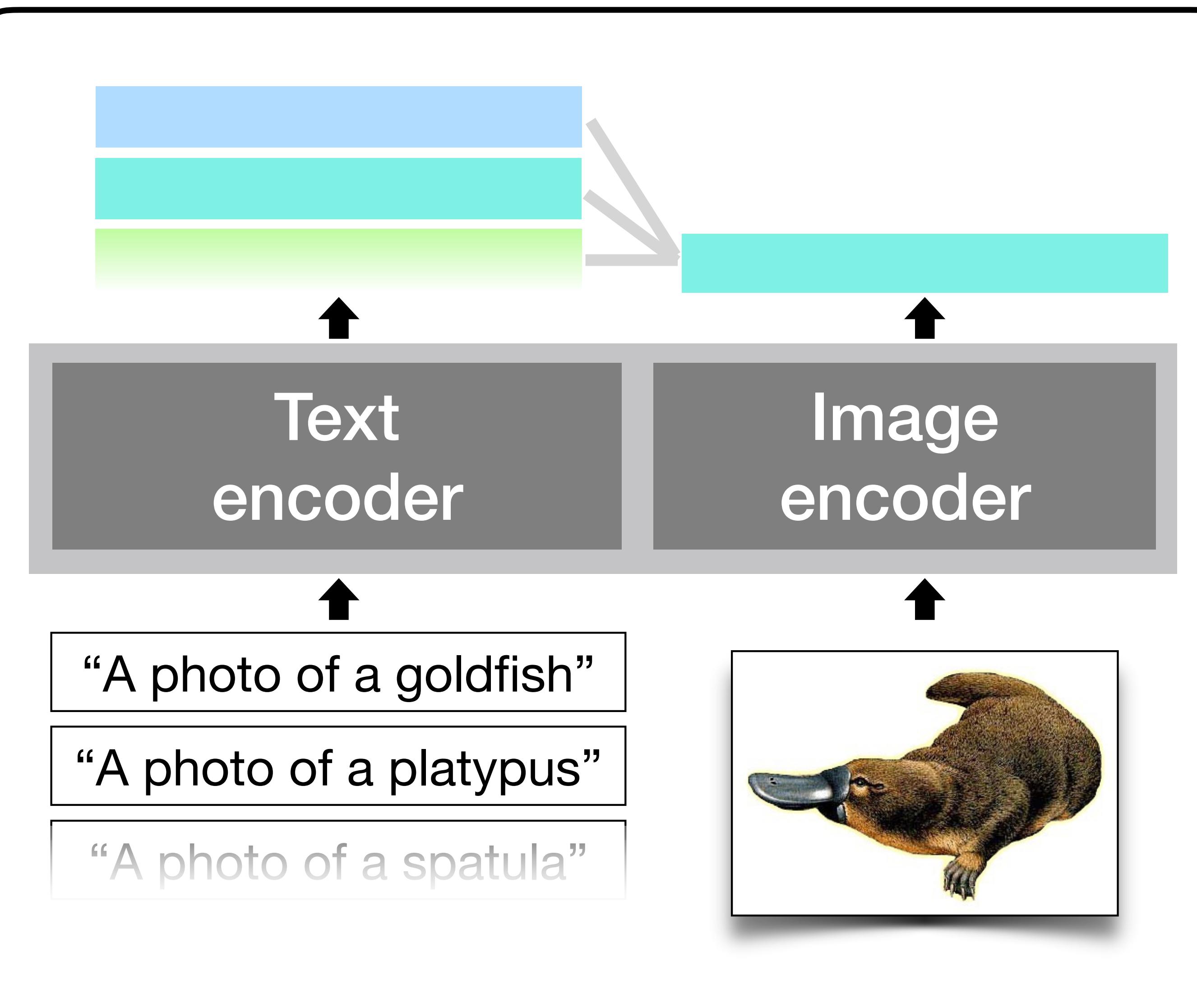
Standard Zero-shot



Wants:

- **Captions with specific visual information**
- 2. **Many hand-written prompts**
- 3. **Contain information about data distribution**

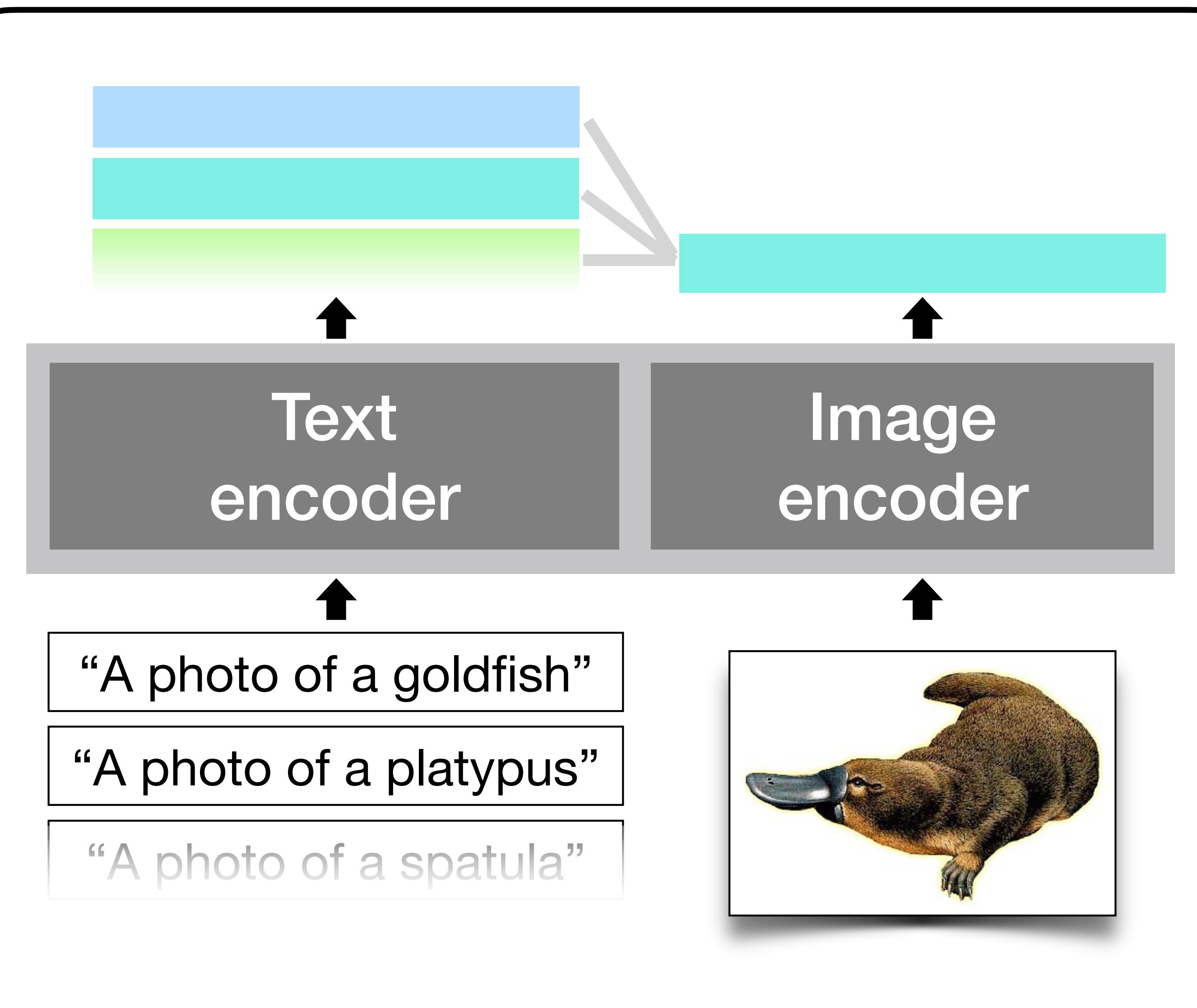
Standard Zero-shot



Wants:

- **Captions with specific visual information**
 - **Fewer handwritten prompts**
- 3. Contain information about data distribution**

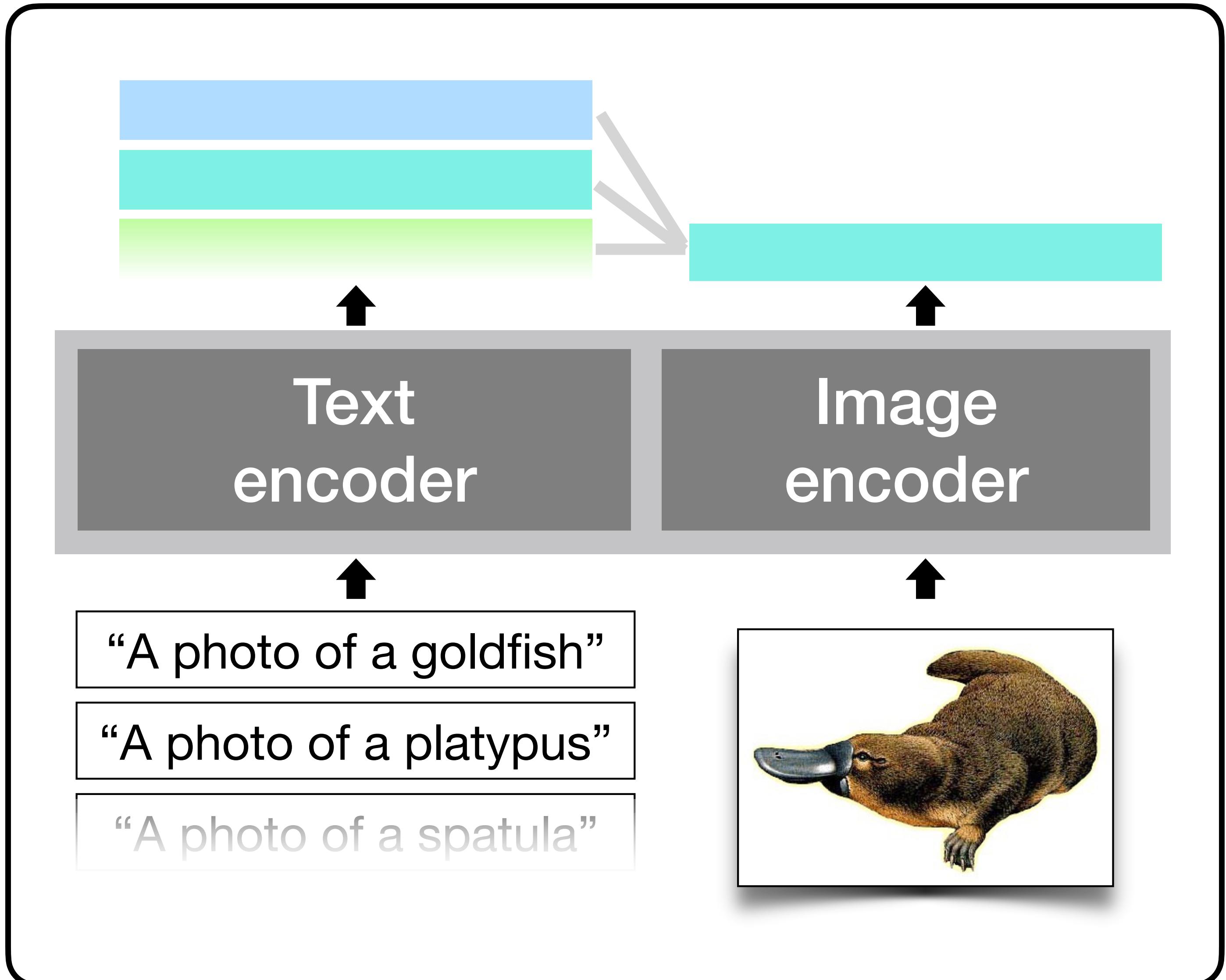
Standard Zero-shot

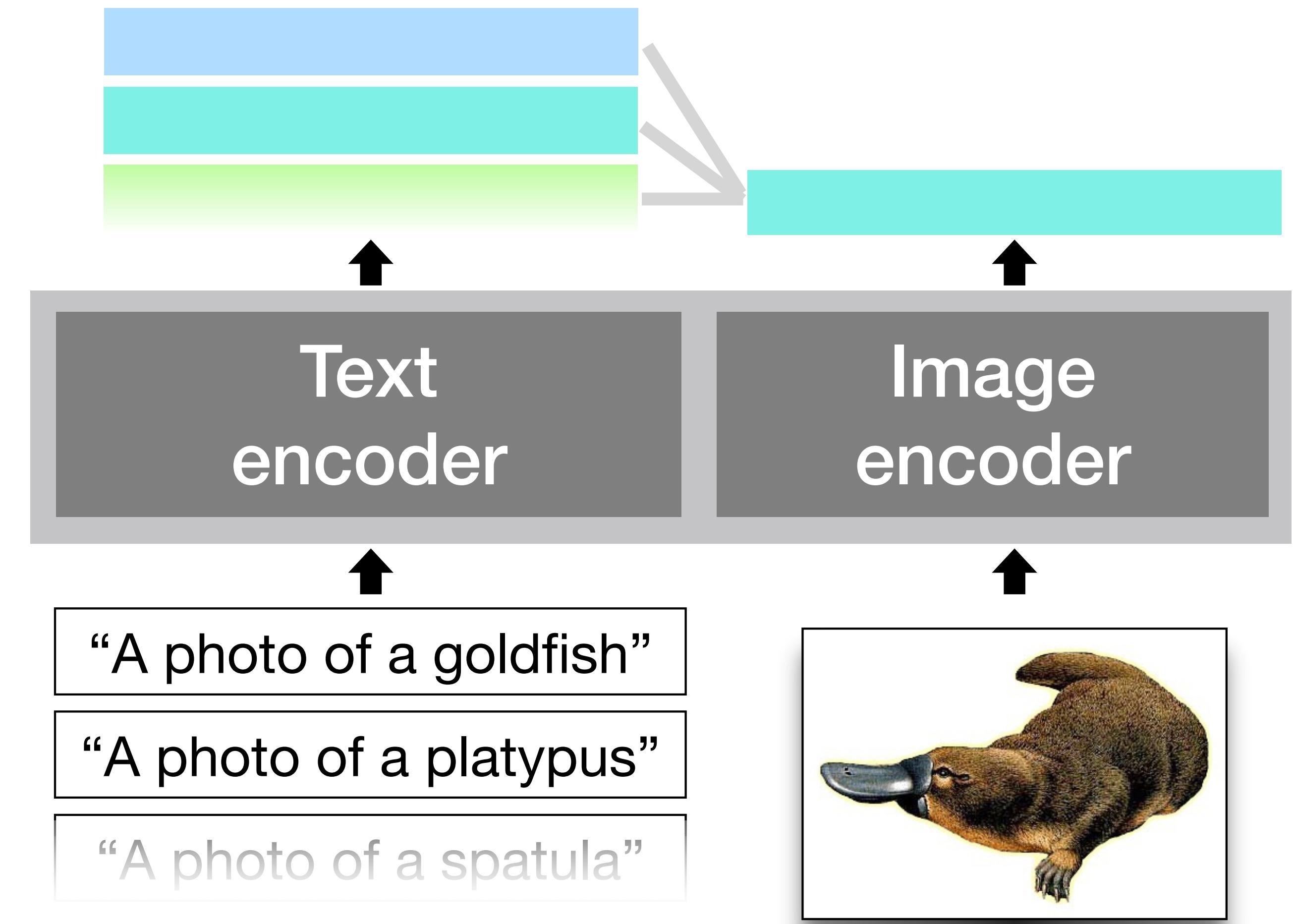


Wants:

- **Captions with specific visual information**
- **Fewer handwritten prompts**
- **No info about data distribution**

Standard Zero-shot

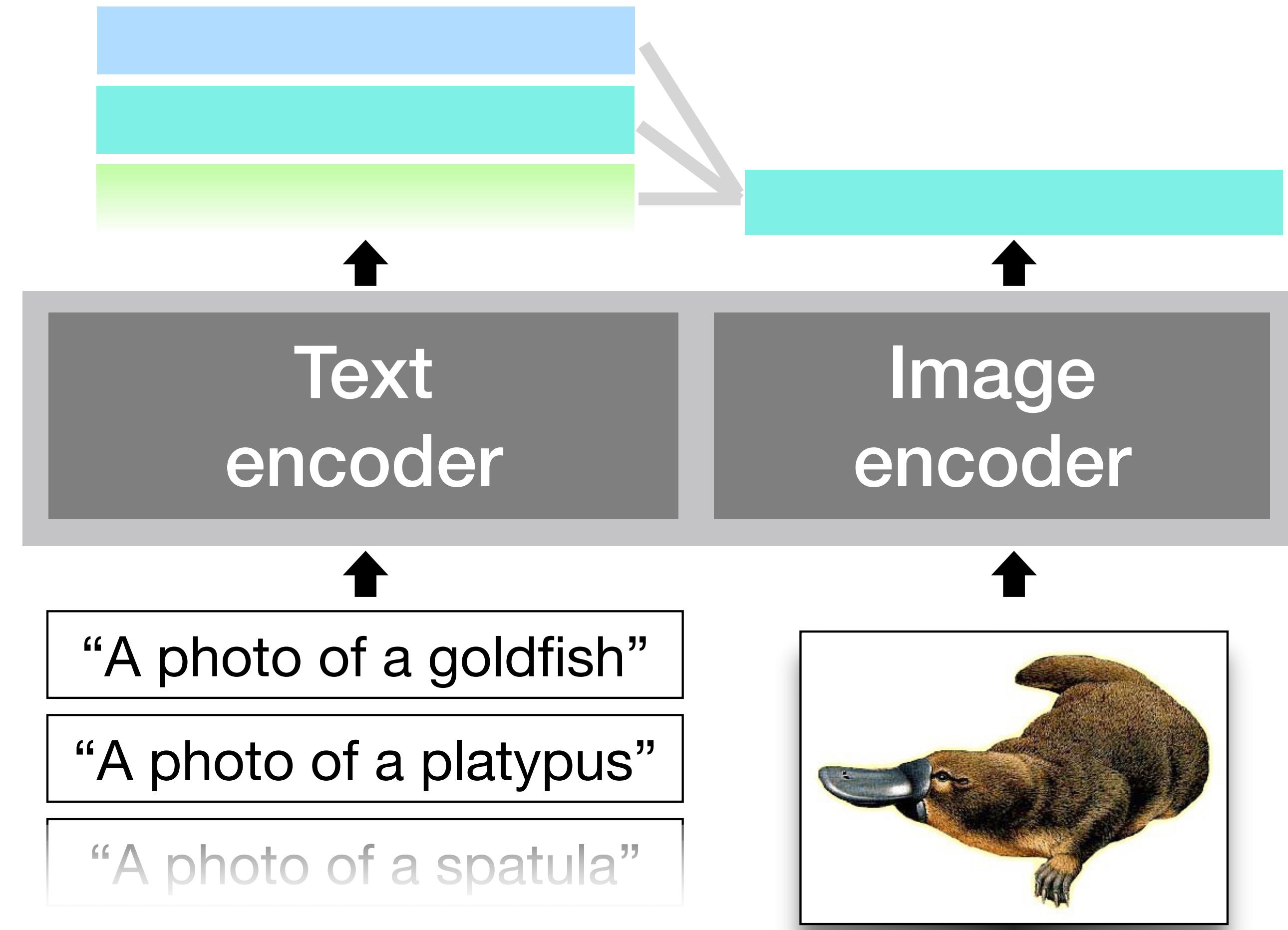


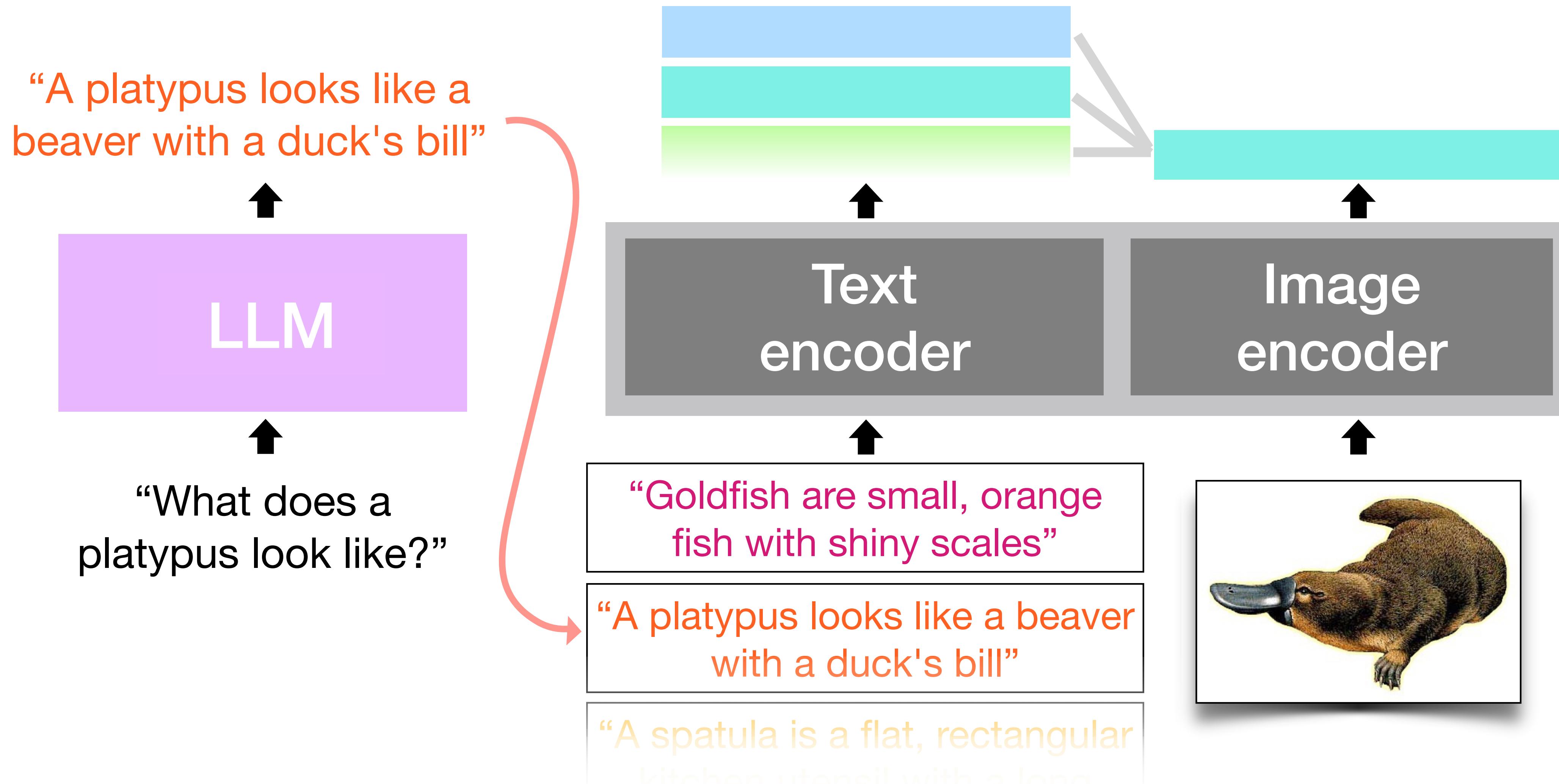


“A platypus looks like a beaver with a duck's bill”

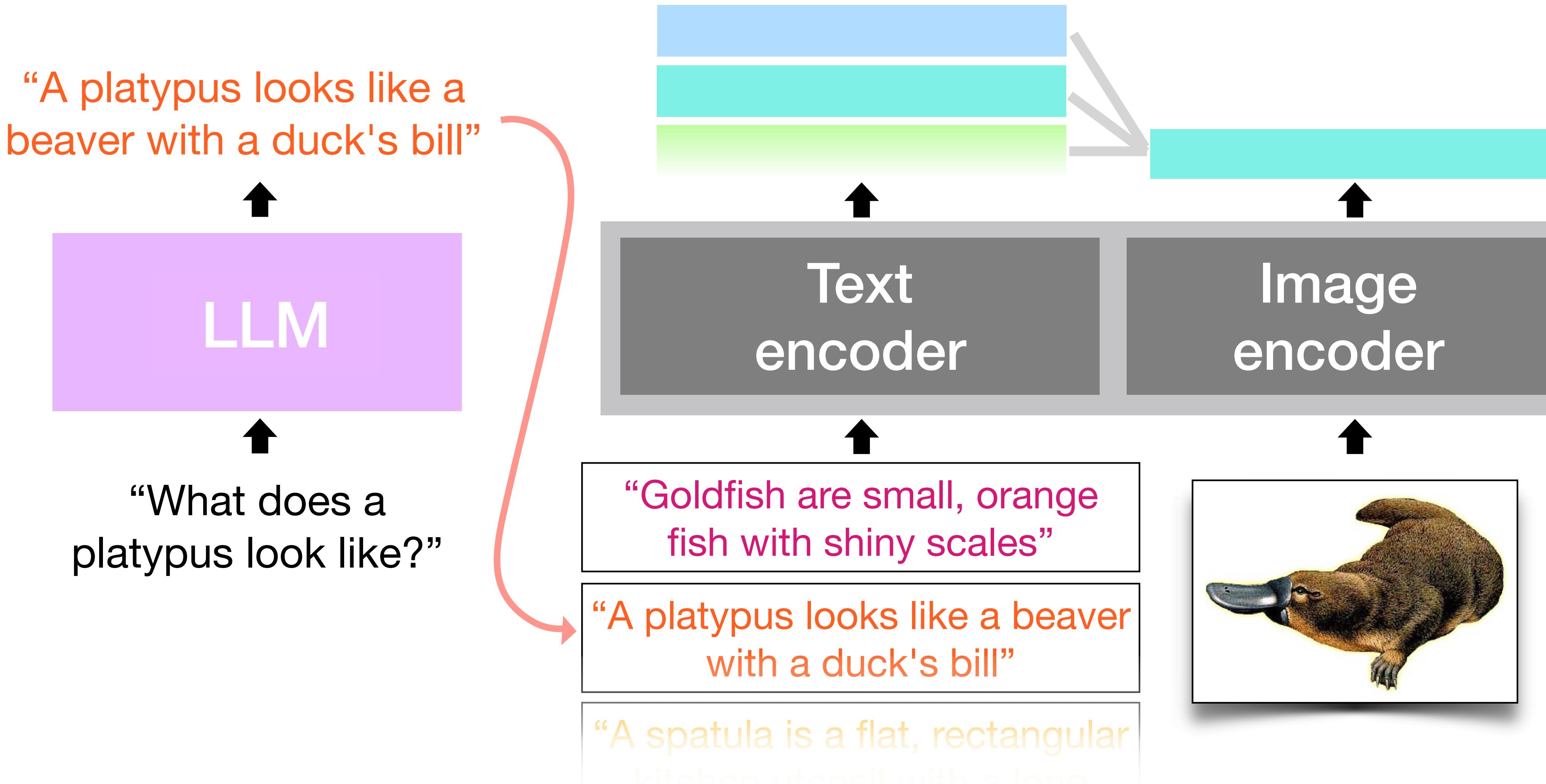
LLM

“What does a platypus look like?”



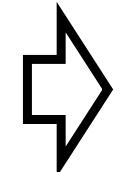


Customized Prompts via Language Models (CuPL)



LLM-prompts:

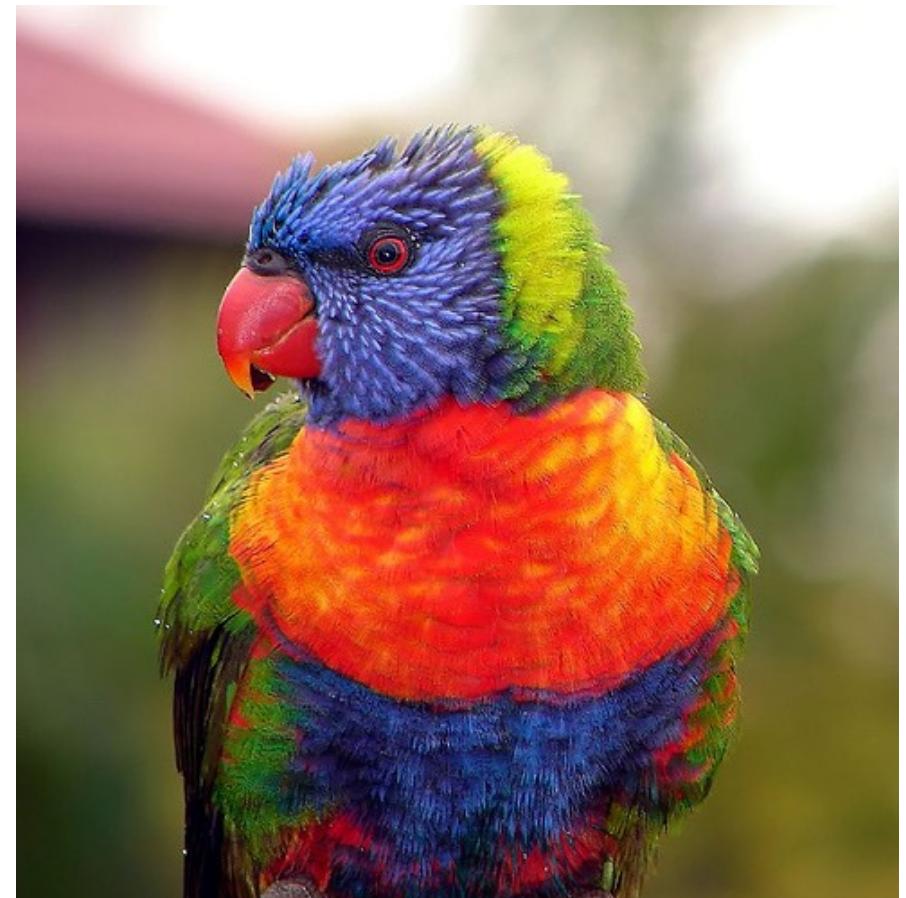
“What does a
{[lorikeet](#), [marimba](#),
[viaduct](#), [papillon](#)}
look like?”



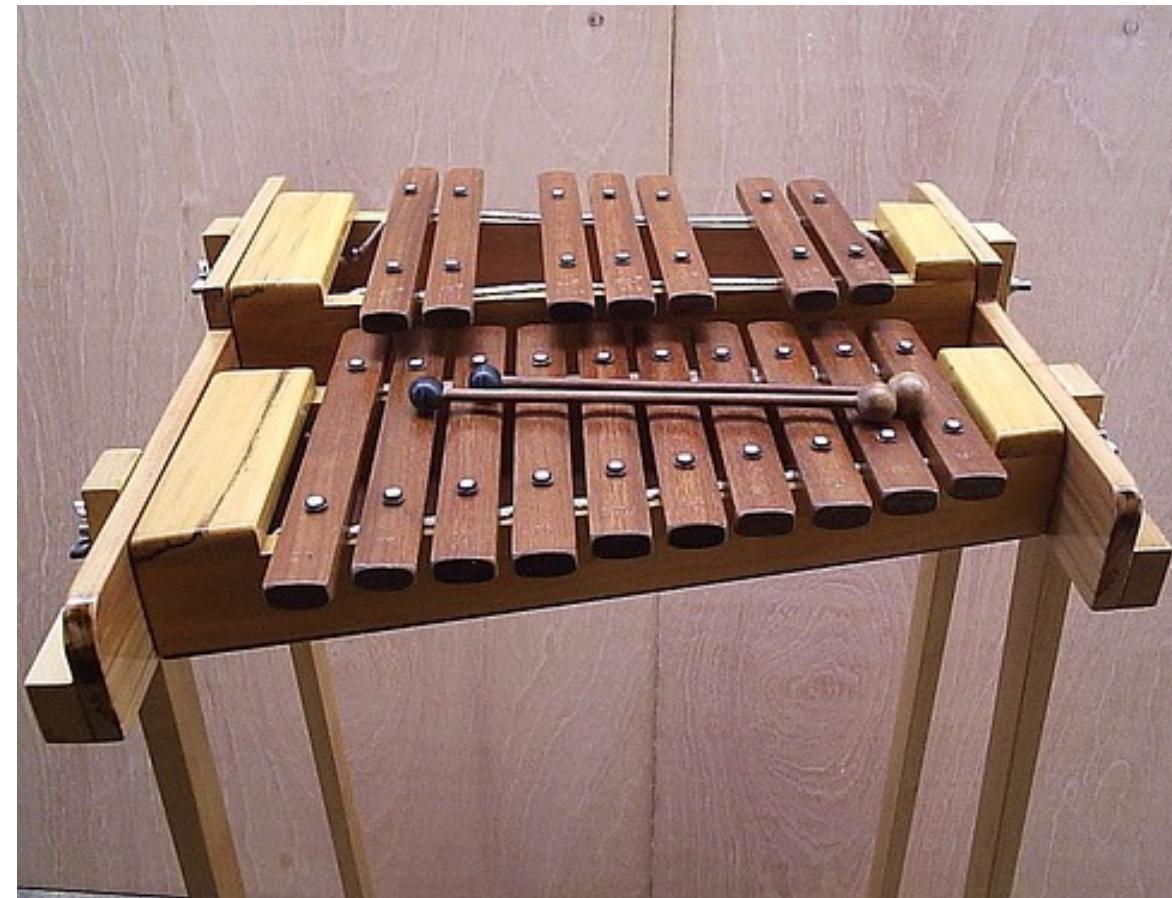
LLM

Image-prompts:

“A [lorikeet](#) is a small to medium-sized parrot with a brightly colored plumage.”
“A [marimba](#) is a large wooden percussion instrument that looks like a xylophone.”
“A [viaduct](#) is a bridge composed of several spans supported by piers or pillars.”
“A [papillon](#) is a small, spaniel-type dog with a long, silky coat and fringed ears.”



Lorikeet



Marimba



Viaduct



Papillon

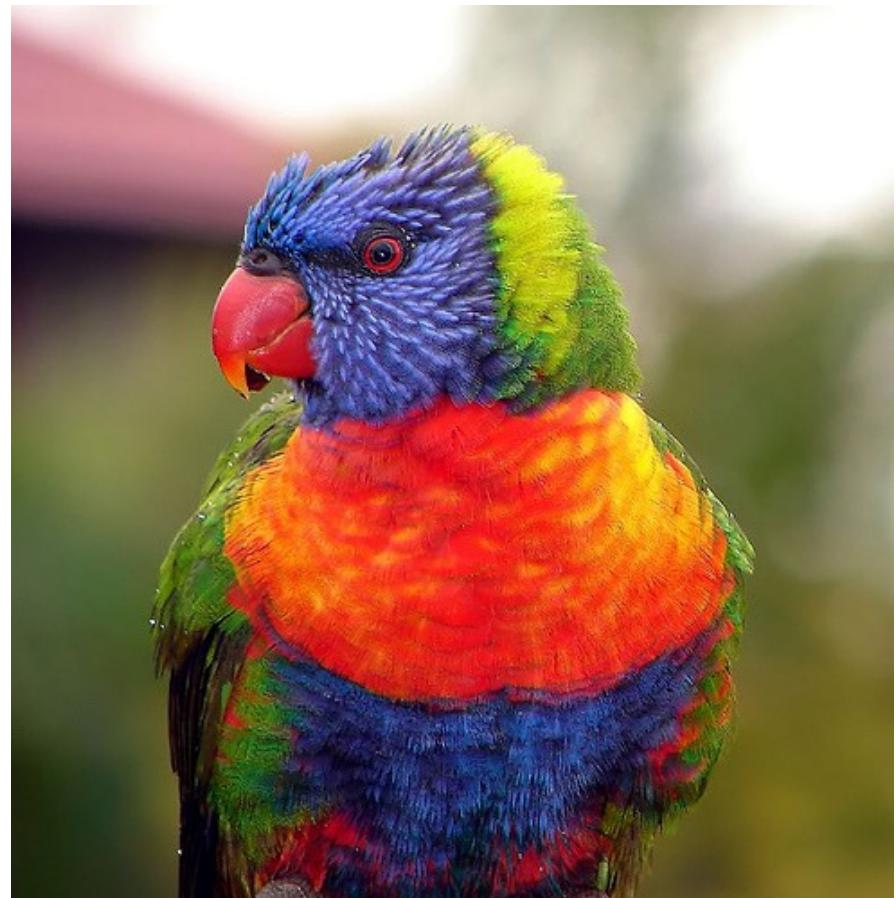
LLM-prompts:

“What does a
{[lorikeet](#), [marimba](#),
[viaduct](#), [papillon](#)}
look like?”

LLM

Image-prompts:

“A [lorikeet](#) is a small to medium-sized parrot with a brightly colored plumage.”
“A [marimba](#) is a large wooden percussion instrument that looks like a xylophone.”
“A [viaduct](#) is a bridge composed of several spans supported by piers or pillars.”
“A [papillon](#) is a small, spaniel-type dog with a long, silky coat and fringed ears.”



Lorikeet



Marimba



Viaduct



Papillon

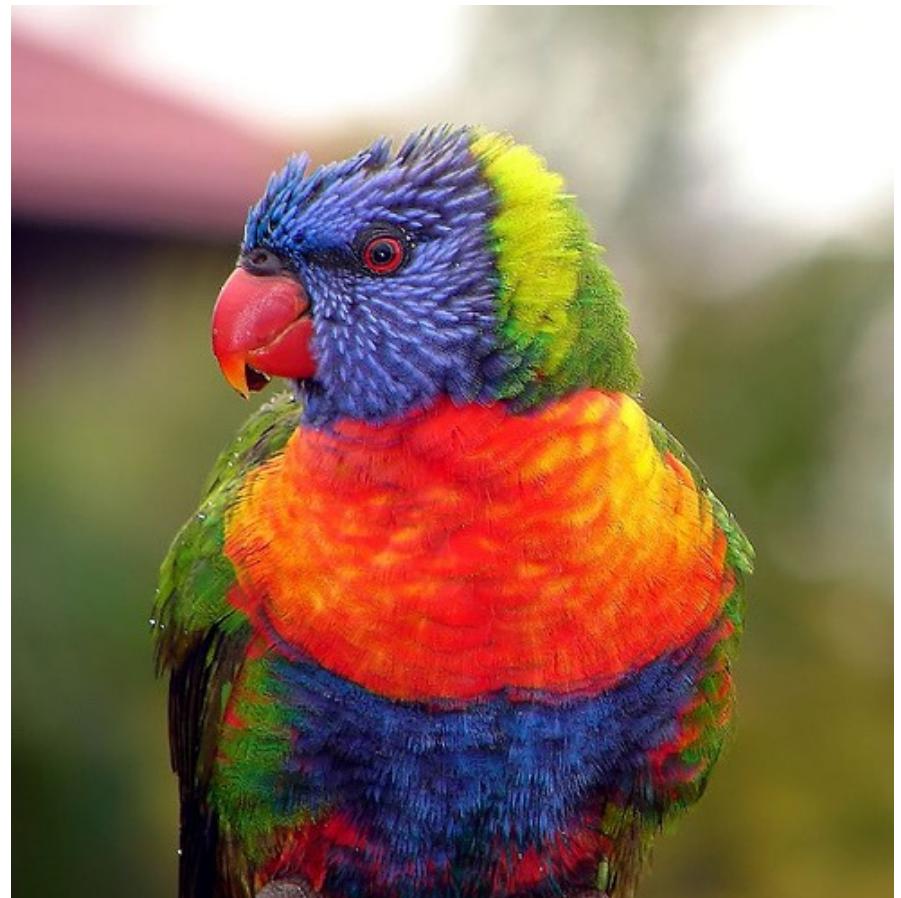
LLM-prompts:

“What does a
{[lorikeet](#), [marimba](#),
[viaduct](#), [papillon](#)}
look like?”

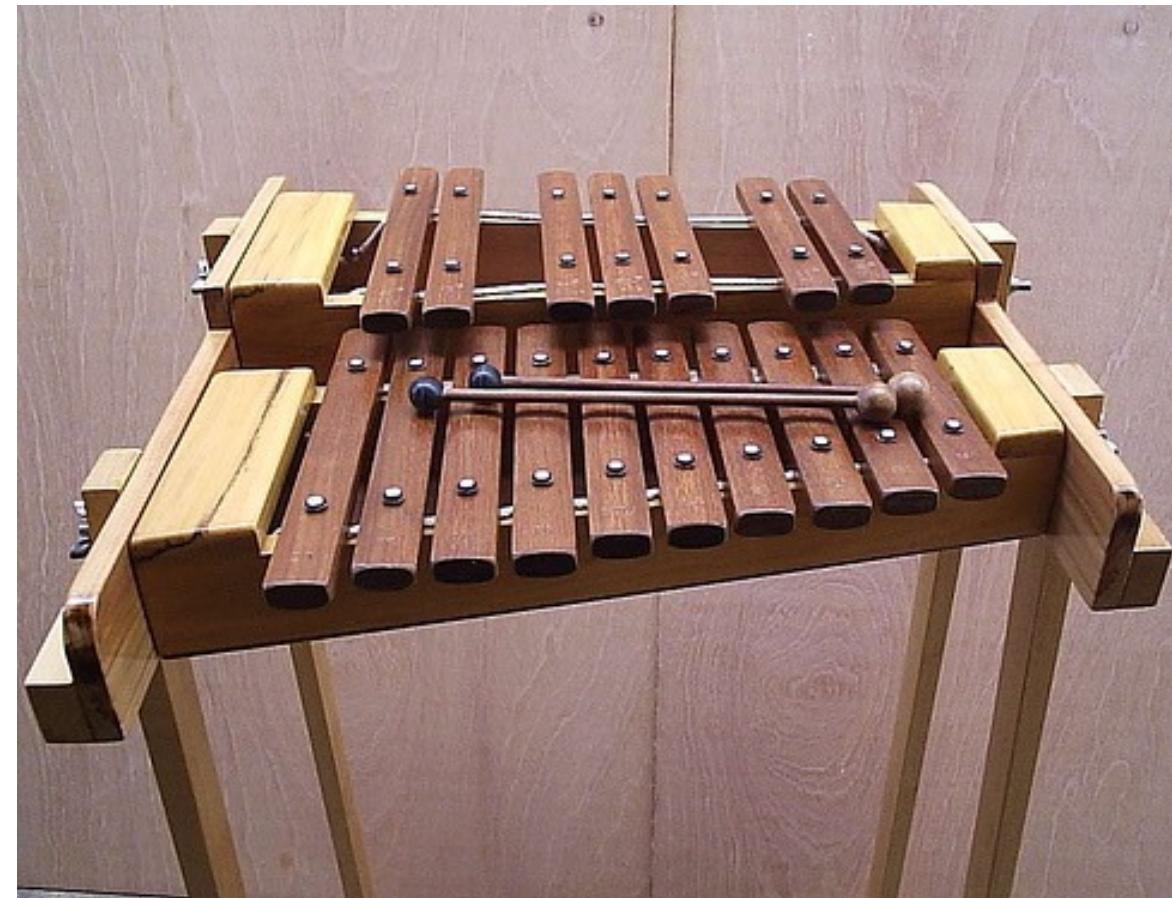
LLM

Image-prompts:

“A [lorikeet](#) is a small to medium-sized parrot with a brightly colored plumage.”
“A [marimba](#) is a large wooden percussion instrument that looks like a xylophone.”
“A [viaduct](#) is a bridge composed of several spans supported by piers or pillars.”
“A [papillon](#) is a small, spaniel-type dog with a long, silky coat and fringed ears.”



Lorikeet



Marimba



Viaduct



Papillon

Standard

ImageNet

a bad photo of a {}.
a photo of many {}.
a sculpture of a {}.
a photo of the hard to see {}.
a low resolution photo of the {}.
a rendering of a {}.
graffiti of a {}.

+ 73 additional prompts

Kinetics-700

a photo of {}.
a photo of a person {}.
a photo of a person using {}.
a photo of a person doing {}.
a photo of a person during {}.
a photo of a person performing {}.
a photo of a person practicing {}.

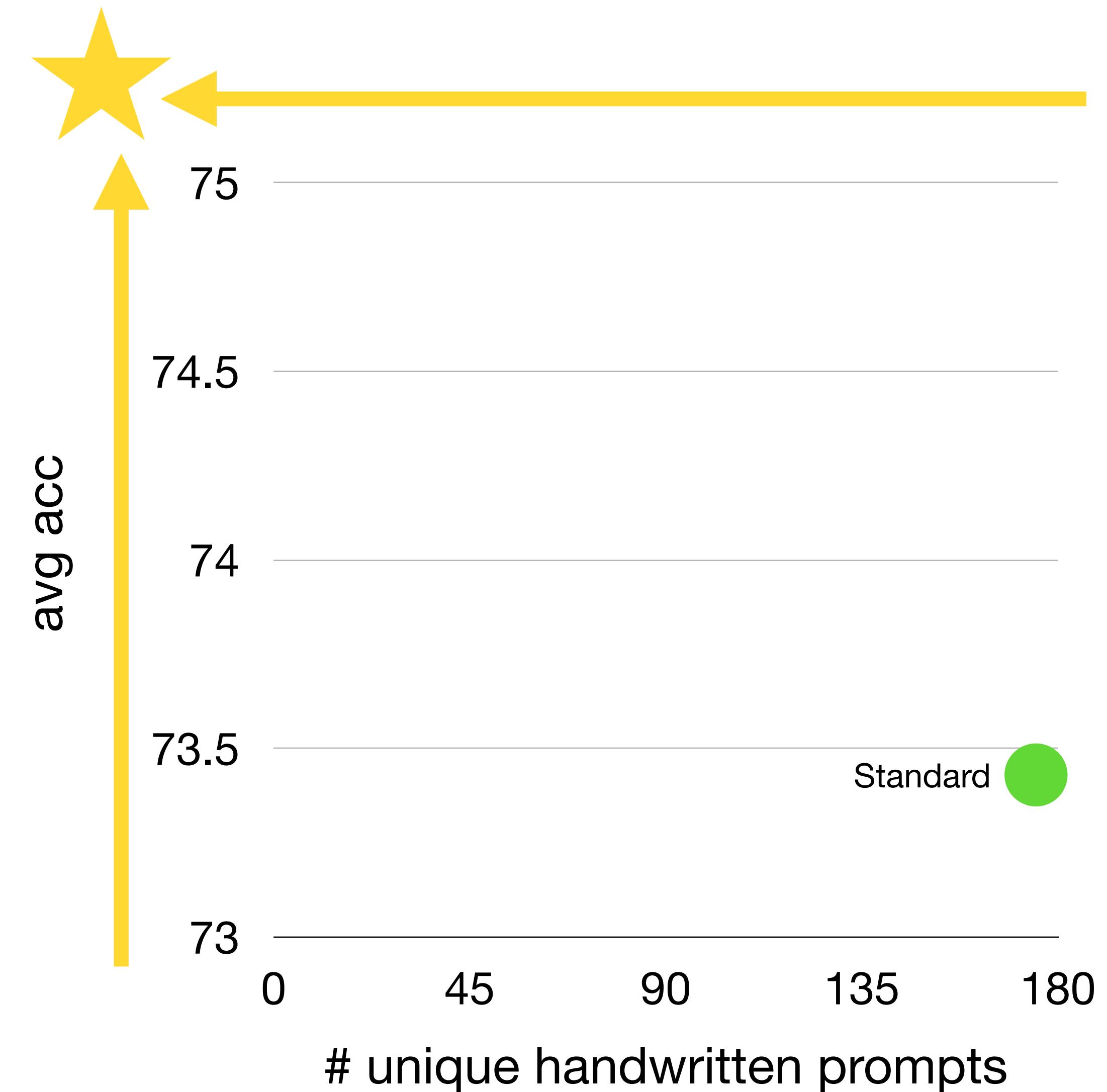
+ 21 Additional prompts

FGVC Aircraft

a photo of a {} a type of aircraft.
a photo of the {} a type of aircraft.

	ImageNet	DTD	Stanford Cars	SUN397	Food101	FGVC Aircraft	Oxford Pets	Caltech101	Flowers 102	UCF101	Kinetics-700	RESISC45	CIFAR-10	CIFAR-100	Birdsnap
std accuracy	75.54	55.20	77.53	69.31	93.08	32.88	93.33	93.24	78.53	77.45	60.07	71.10	95.59	78.26	50.43
# hw	80	8	8	2	1	2	1	34	1	48	28	18	18	18	1

Accuracy vs # Prompts



Standard

ImageNet

a bad photo of a {}.
a photo of many {}.
a sculpture of a {}.
a photo of the hard to see {}.
a low resolution photo of the {}.
a rendering of a {}.
graffiti of a {}.

+ 73 additional prompts

Kinetics-700

a photo of {}.
a photo of a person {}.
a photo of a person using {}.
a photo of a person doing {}.
a photo of a person during {}.
a photo of a person performing {}.
a photo of a person practicing {}.

+ 21 Additional prompts

FGVC Aircraft

a photo of a {} a type of aircraft.
a photo of the {} a type of aircraft.

**CuPL
Full**

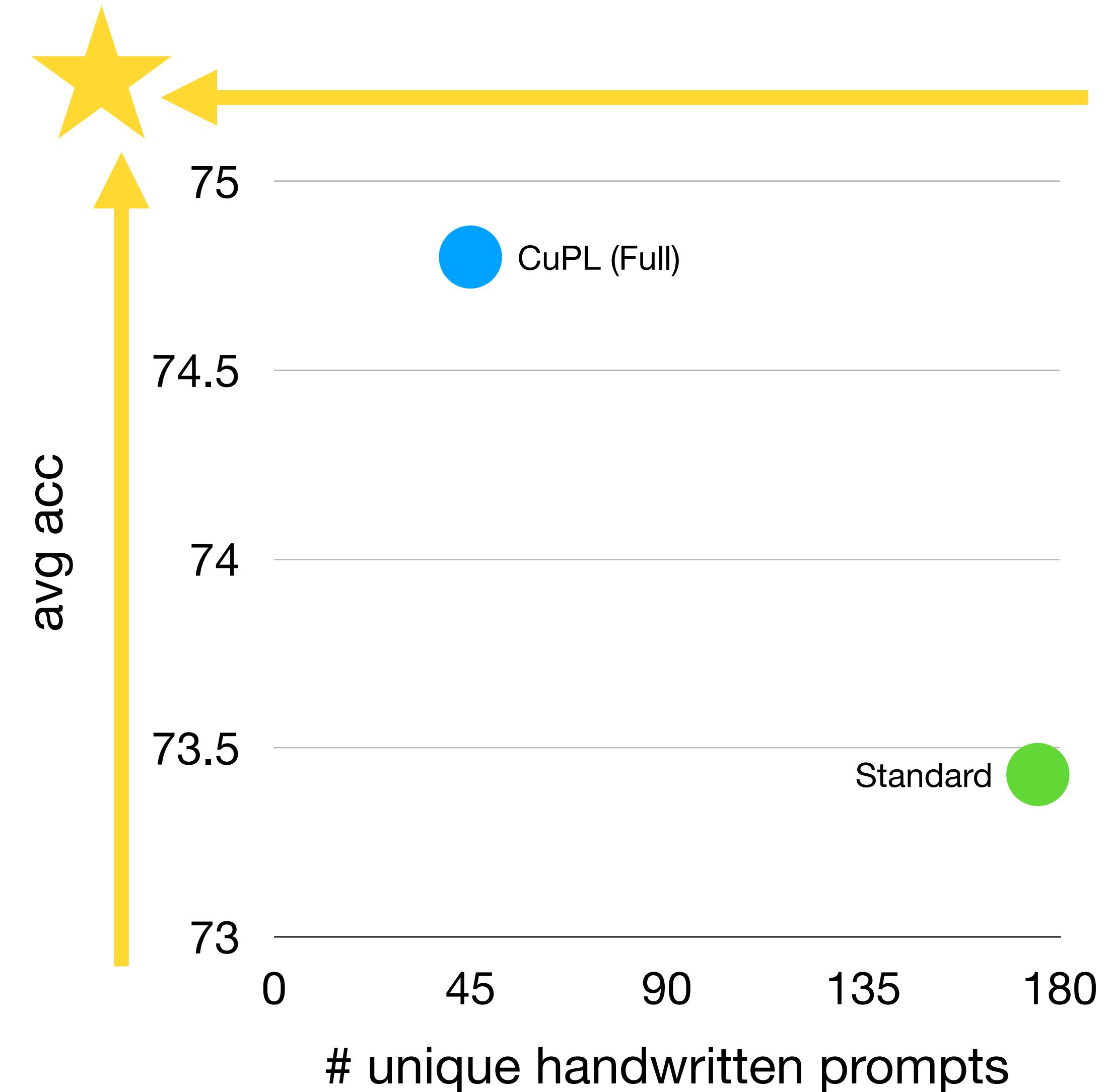
Describe what a(n) {} looks like
How can you identify a(n) {}?
What does a(n) {} look like?
Describe an image from the internet of a(n) {}
A caption of an image of a(n) {}:

Describe the action “{}”
What does a person {} look like?
What does the act of {} look like?
Describe “{}”

Describe a(n) {} aircraft
Describe the {} aircraft

	ImageNet	DTD	Stanford Cars	SUN397	Food101	FGVC Aircraft	Oxford Pets	Caltech101	Flowers 102	UCF101	Kinetics-700	RESISC45	CIFAR-10	CIFAR-100	Birdsnap
std accuracy	75.54	55.20	77.53	69.31	93.08	32.88	93.33	93.24	78.53	77.45	60.07	71.10	95.59	78.26	50.43
# hw	80	8	8	2	1	2	1	34	1	48	28	18	18	18	1
CuPL (full)	76.69	61.70	77.63	73.31	93.36	36.11	93.81	93.45	79.67	78.36	60.63	71.69	95.84	78.57	51.11
Δ std	+1.15	+6.50	+0.10	+4.00	+0.28	+3.23	+0.48	+0.21	+1.14	+0.91	+0.56	+0.59	+0.25	+0.31	+0.63
# hw	5	6	9	3	3	2	2	3	2	5	4	5	3	4	3

Accuracy vs # Prompts



Standard

ImageNet

a bad photo of a {}.
a photo of many {}.
a sculpture of a {}.
a photo of the hard to see {}.
a low resolution photo of the {}.
a rendering of a {}.
graffiti of a {}.

+ 73 additional prompts

Kinetics-700

a photo of {}.
a photo of a person {}.
a photo of a person using {}.
a photo of a person doing {}.
a photo of a person during {}.
a photo of a person performing {}.
a photo of a person practicing {}.

+ 21 Additional prompts

FGVC Aircraft

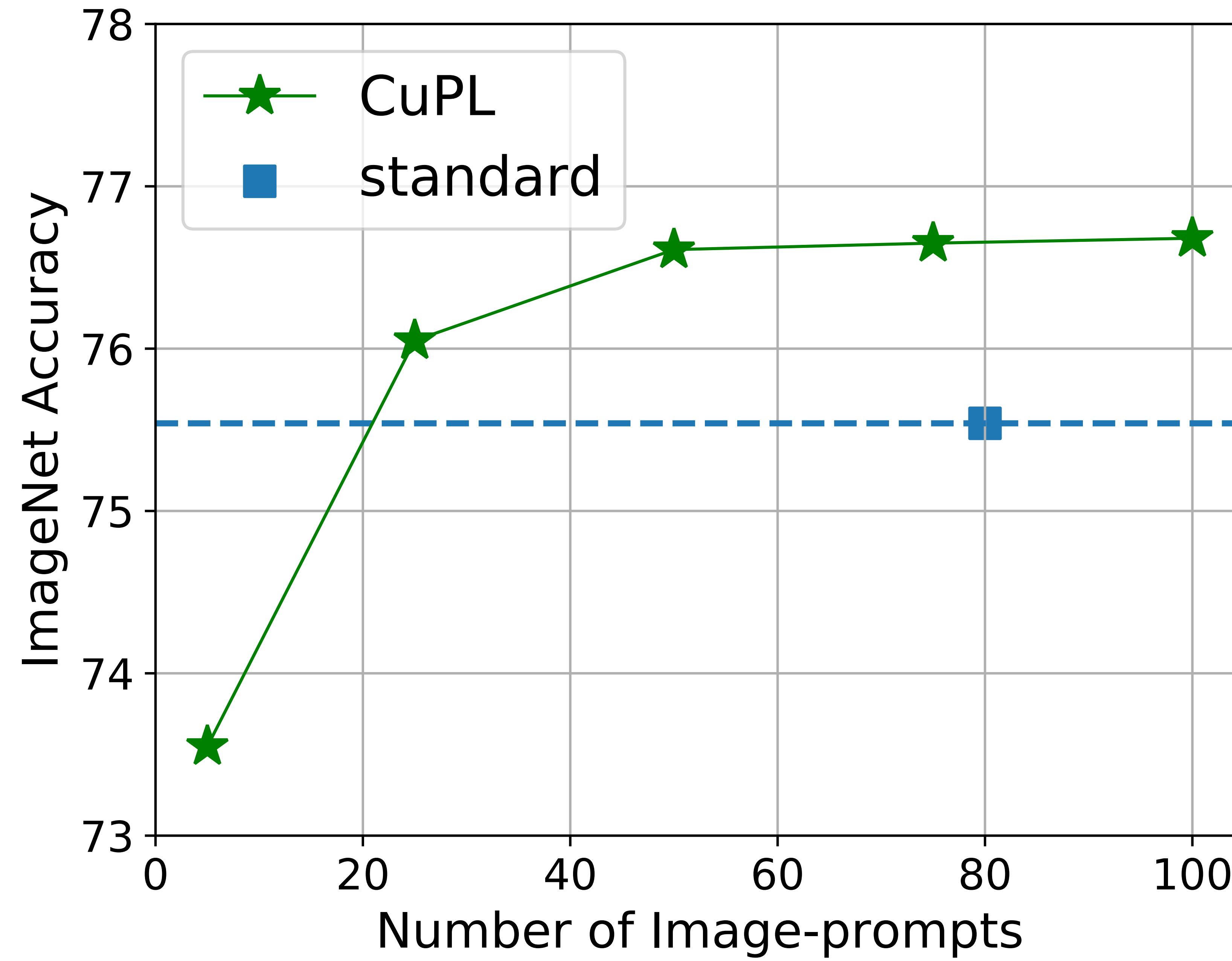
a photo of a {} a type of aircraft.
a photo of the {} a type of aircraft.

**CuPL
Full**

Describe what a(n) {} looks like
How can you identify a(n) {}?
What does a(n) {} look like?
Describe an image from the internet of a(n) {}
A caption of an image of a(n) {}:

Describe the action “{}”
What does a person {} look like?
What does the act of {} look like?
Describe “{}”

Describe a(n) {} aircraft
Describe the {} aircraft



Standard

ImageNet

a bad photo of a {}.
a photo of many {}.
a sculpture of a {}.
a photo of the hard to see {}.
a low resolution photo of the {}.
a rendering of a {}.
graffiti of a {}.

+ 73 additional prompts

Kinetics-700

a photo of {}.
a photo of a person {}.
a photo of a person using {}.
a photo of a person doing {}.
a photo of a person during {}.
a photo of a person performing {}.
a photo of a person practicing {}.

+ 21 Additional prompts

FGVC Aircraft

a photo of a {} a type of aircraft.
a photo of the {} a type of aircraft.

**CuPL
Full**

Describe what a(n) {} looks like
How can you identify a(n) {}?
What does a(n) {} look like?
Describe an image from the internet of a(n) {}
A caption of an image of a(n) {}:

Describe the action “{}”
What does a person {} look like?
What does the act of {} look like?
Describe “{}”

Describe a(n) {} aircraft
Describe the {} aircraft

Standard

ImageNet

a bad photo of a {}.
a photo of many {}.
a sculpture of a {}.
a photo of the hard to see {}.
a low resolution photo of the {}.
a rendering of a {}.
graffiti of a {}.

+ 73 additional prompts

Kinetics-700

a photo of {}.
a photo of a person {}.
a photo of a person using {}.
a photo of a person doing {}.
a photo of a person during {}.
a photo of a person performing {}.
a photo of a person practicing {}.

+ 21 Additional prompts

FGVC Aircraft

a photo of a {} a type of aircraft.
a photo of the {} a type of aircraft.

CuPL Full

Describe what a(n) {} looks like
How can you identify a(n) {}?
What does a(n) {} look like?
Describe an image from the internet of a(n) {}
A caption of an image of a(n) {}:

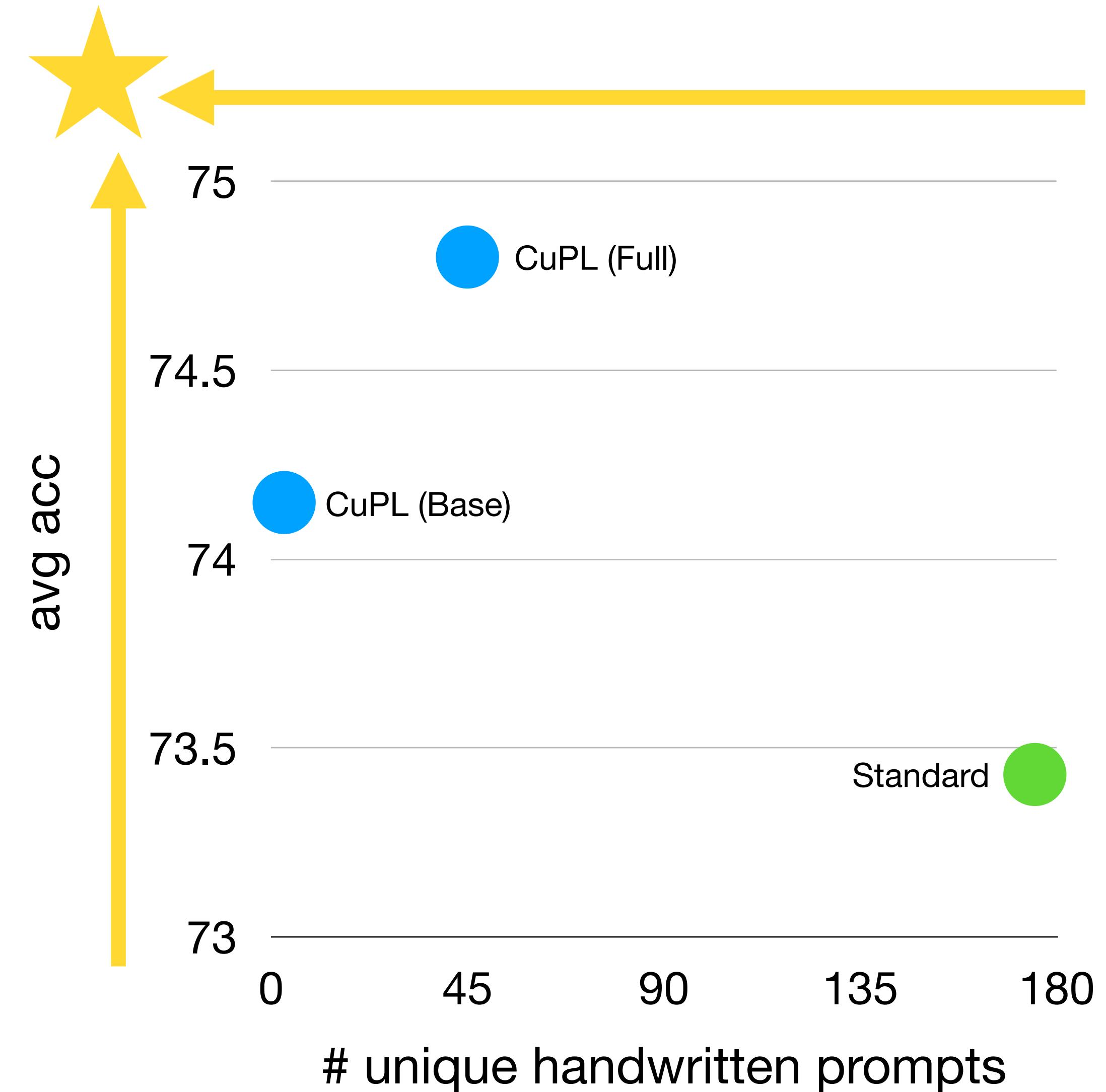
Describe the action “{}”
What does a person {} look like?
What does the act of {} look like?
Describe “{}”

Describe a(n) {} aircraft
Describe the {} aircraft

CuPL Base

Describe what a/the {dataset type} {classname} looks like:
Describe a/the {dataset type} {classname}:
What are the identifying characteristics of a/the {dataset type} {classname}?

Accuracy vs # Prompts



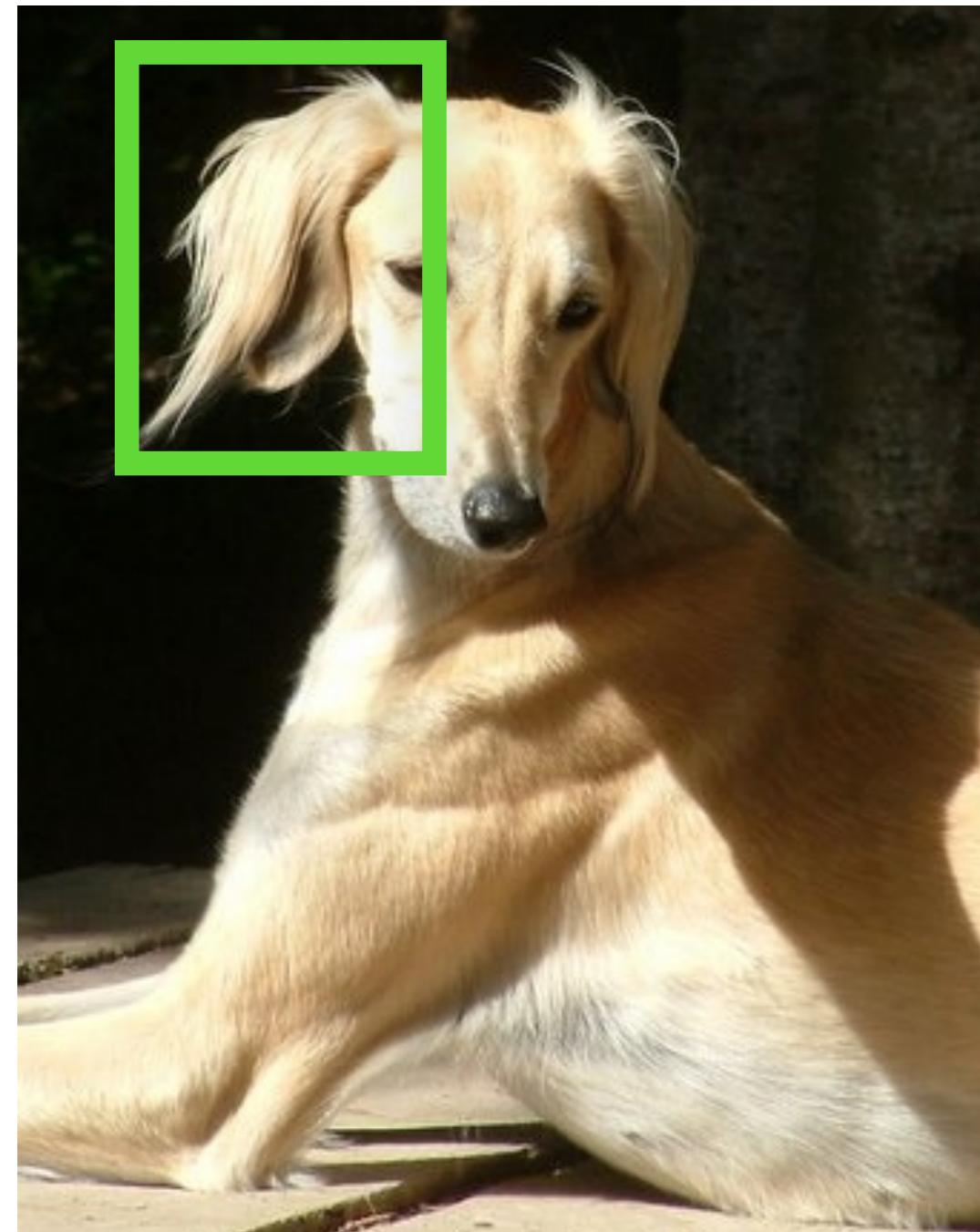
Why do descriptions help?

Why do descriptions help?



The easiest way to identify a Saluki is by its iconic long, silky ears.

Why do descriptions help?



The easiest way to identify a Saluki is by its iconic long, silky ears.

Promontory Prompts



A photo of a
promontory

Cliff Prompts



A photo of a cliff

Standard

Promontory Prompts



A photo of a
promontory

Cliff Prompts



A photo of a cliff

CuPL

Promontory Prompts



A promontory is a
landform that protrudes
into a body of water.

Cliff Prompts



A cliff is a high, steep
rock face or slope.

Standard

Promontory Prompts



A photo of a
promontory

Cliff Prompts



A photo of a cliff

CuPL

Promontory Prompts



A promontory is a
landform that protrudes
into a body of water.

Cliff Prompts



A cliff is a high, steep
rock face or slope.

Standard

Promontory Prompts



A photo of a promontory

Cliff Prompts



A photo of a cliff

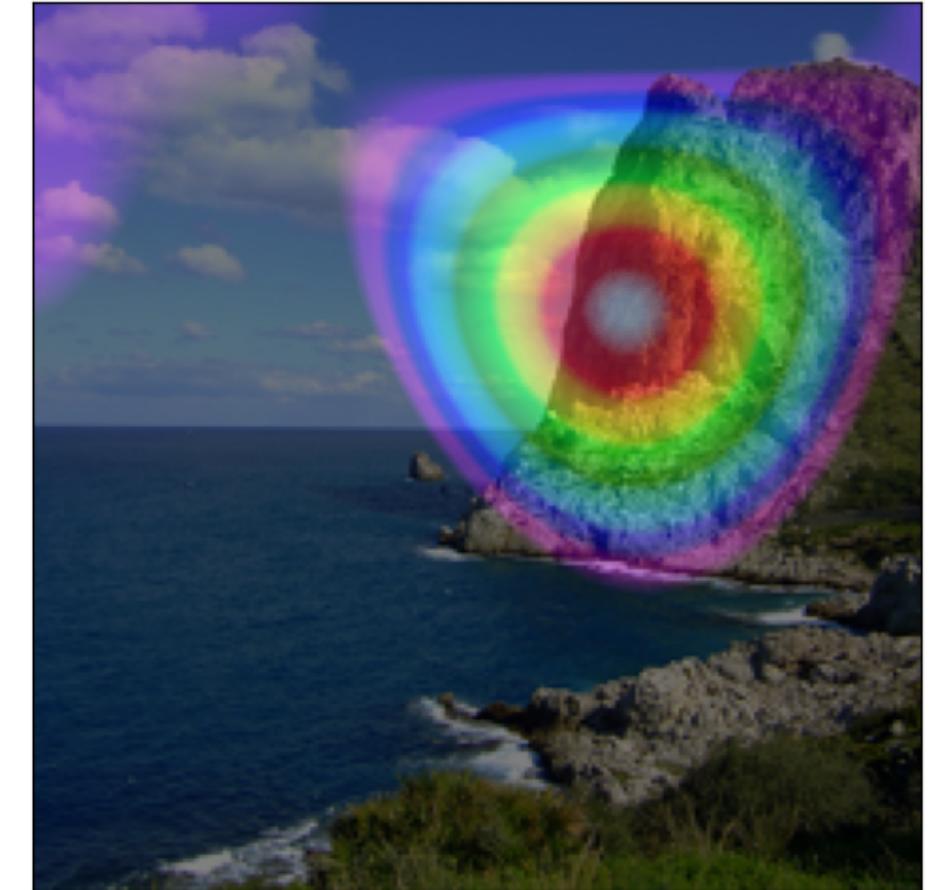
CuPL

Promontory Prompts



A promontory is a landform that protrudes into a body of water.

Cliff Prompts



A cliff is a high, steep rock face or slope.

Standard

Promontory Prompts



A photo of a
promontory

Prediction: Cliff



Cliff Prompts



A photo of a cliff

Promontory Prompts



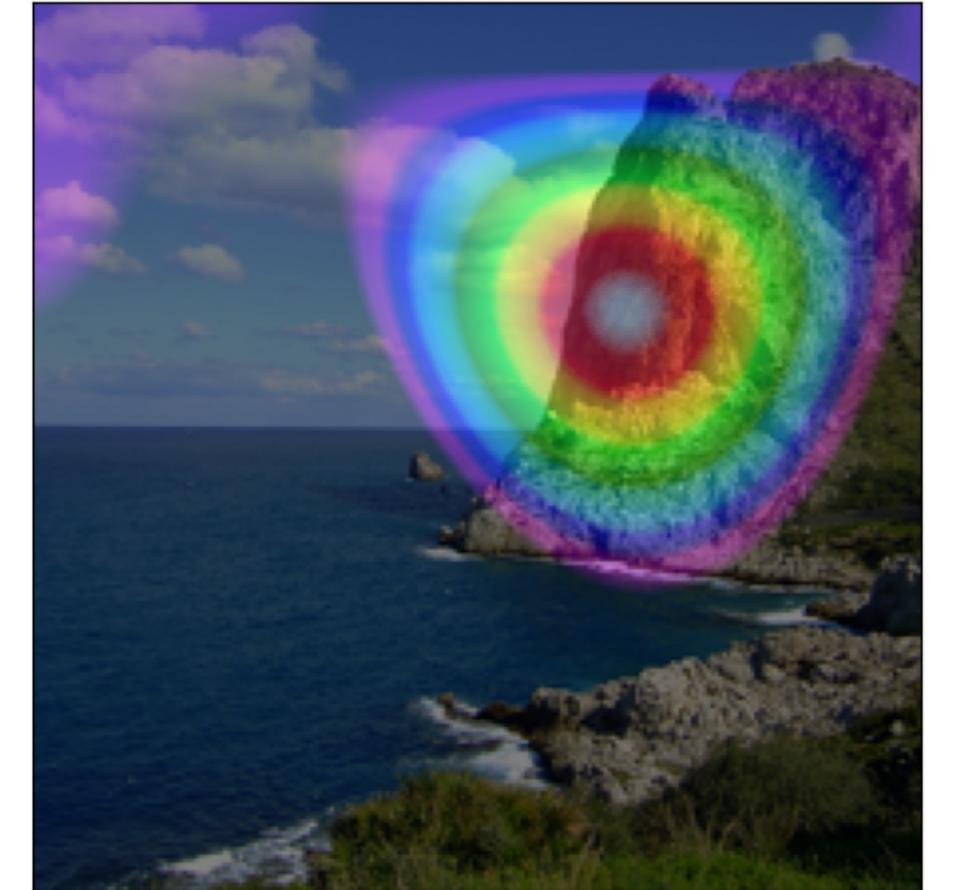
A promontory is a
landform that protrudes
into a body of water.

Prediction: Promontory



CuPL

Cliff Prompts



A cliff is a high, steep
rock face or slope.

Standard

Tree Frog Prompts



A photo of a tree frog

Tailed Frog Prompts



A photo of a tailed frog

CuPL

Tree Frog Prompts



A tree frog looks like a small frog with large eyes.

Tailed Frog Prompts



The tailed frog is a small frog that is found in North America.

Standard

Tree Frog Prompts



A photo of a tree frog

Tailed Frog Prompts



A photo of a tailed frog

CuPL

Tree Frog Prompts



A tree frog looks like a small frog with large eyes.

Tailed Frog Prompts



The tailed frog is a small frog that is found in North America.

Standard

Tree Frog Prompts



A photo of a tree frog

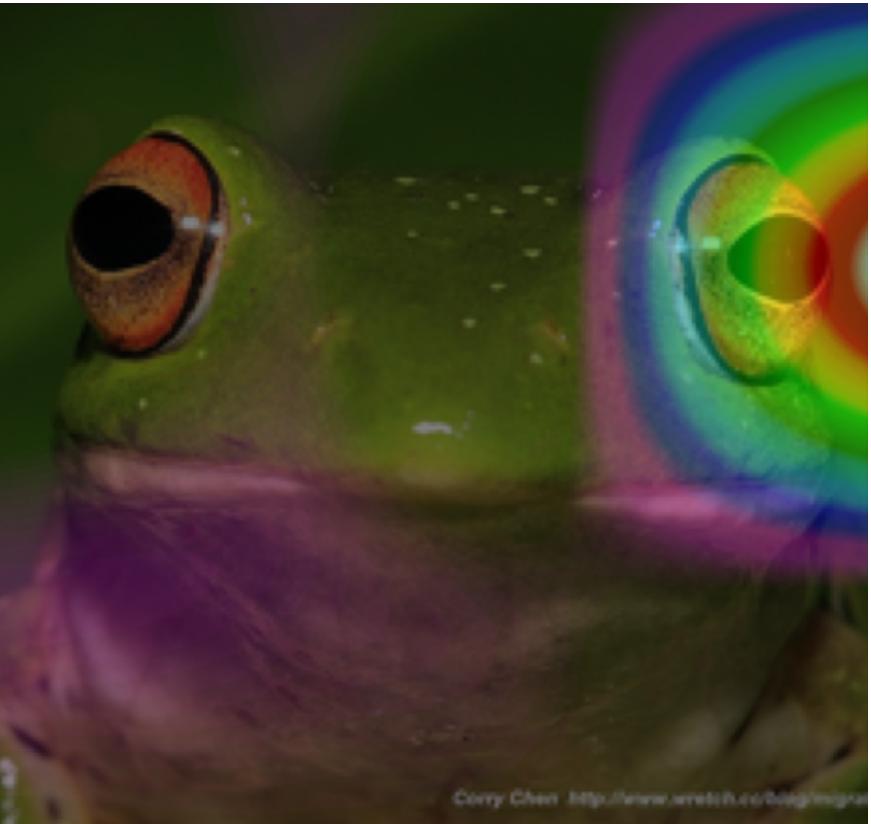
Tailed Frog Prompts



A photo of a tailed frog

CuPL

Tree Frog Prompts



A tree frog looks like a small frog with large eyes.

Tailed Frog Prompts



The tailed frog is a small frog that is found in North America.

Standard

Tree Frog Prompts



A photo of a tree frog

Tailed Frog Prompts



A photo of a tailed frog

CuPL

Tree Frog Prompts



A tree frog looks like a small frog with large eyes.

Tailed Frog Prompts



The tailed frog is a small frog that is found in North America.

Prediction: Tailed Frog



Prediction: Tree Frog



CuPL in the wild

CuPL in the wild

**REACHING 80% ZERO-SHOT
ACCURACY WITH OPENCLIP: VIT-G/14
TRAINED ON LAION-2B**

Model name	Batch size	Samples seen	Text Params	Image params	ImageNet top1	Mscoco image retrieval at 5	Flickr30k image retrieval at 5
OpenAI CLIP L/14	32k	13B	123.65M	303.97M	75.4%	61.0%	87.0%
OpenCLIP H/14	79k	32B (16 epochs of laion2B)	354.0M	632.08M	78.0%	73.4%	94%
OpenCLIP G/14	160k	32B +unmasked fine-tune (details below)	694.7M	1844.9M	80.1%*	74.9%	94.9%
CoCa	66k	33B	1100M	1000M	86.3%**	74.2	95.7

* When using CuPL prompts instead of the standard prompts from OpenAI, the zero-shot accuracy is 80.3%.

Ilharco, Gabriel, et al.
"OpenClip". 2021.

CuPL in the wild

**REACHING 80% ZERO-SHOT
ACCURACY WITH OPENCLIP: VIT-G/14
TRAINED ON LAION-2B**

Model name	Batch size	Samples seen	80.1%*		ImageNet top1	Mscoco image retrieval at 5	Flickr30k image retrieval at 5
OpenAI CLIP L/14	32k	13B	354.0M	632.08M	75.4%	61.0%	87.0%
OpenCLIP H/14	79k	32B (16 epochs of laion2B)			78.0%	73.4%	94%
OpenCLIP G/14	160k	32B +unmasked fine-tune (details below)	694.7M	1844.9M	80.1%*	74.9%	94.9%
CoCa	66k	33B	1100M	1000M	86.3%**	74.2	95.7

* When using CuPL prompts instead of the standard prompts from OpenAI, the zero-shot accuracy is 80.3%.

Ilharco, Gabriel, et al.
"OpenClip". 2021.

CuPL in the wild

**REACHING 80% ZERO-SHOT
ACCURACY WITH OPENCLIP: VIT-G/14
TRAINED ON LAION-2B**

Model name	Batch size	Samples seen	80.1%*		ImageNet top1	Mscoco image retrieval at 5	Flickr30k image retrieval at 5
OpenAI CLIP L/14	32k	13B	354.0M	632.08M	75.4%	61.0%	87.0%
OpenCLIP H/14	79k	32B (16 epochs of laion2B)			78.0%	73.4%	94%
		32B					

* When using CuPL prompts instead of the standard prompts from OpenAI, the zero-shot accuracy is 80.3%.

		(details below)					
CoCa	66k	33B	1100M	1000M	86.3%**	74.2	95.7

* When using CuPL prompts instead of the standard prompts from OpenAI, the zero-shot accuracy is 80.3%.

Ilharco, Gabriel, et al.
"OpenClip". 2021.

CuPL in the wild

Neural Priming for Sample-Efficient Adaptation

Matthew Wallingford^{*†}, Vivek Ramanujan^{*†}, Alex Fang[†], Aditya Kusupati[†],
Roozbeh Mottaghi[†], Aniruddha Kembhavi[†], Ludwig Schmidt[†], Ali Farhadi[†]
[†]University of Washington, [°]Allen Institute for AI
{mcw244,ramany}@cs.washington.edu

Abstract

We propose Neural Priming, a technique for adapting large pretrained models to distribution shifts and downstream tasks given few or no labeled examples. Presented with class names or unlabeled test samples, Neural Priming enables the model to recall and conditions its parameters on relevant data seen throughout pretraining, thereby priming it for the test distribution. Neural Priming can be performed at test time in even for pretraining datasets as large as LAION-2B. Performing lightweight updates on the recalled data significantly improves accuracy across a variety of distribution shift and transfer learning benchmarks. Concretely, in the zero-shot setting, we see a 2.45% improvement in accuracy on ImageNet and 3.81% accuracy improvement on average across standard transfer learning benchmarks. Further, using our test time inference scheme, we see a 1.41% accuracy improvement on ImageNetV2. These results demonstrate the effectiveness of Neural Priming in addressing the common challenge of limited labeled data and changing distributions. Code is available at github.com/RAIVNLab/neural-priming.

	ImageNet	Stanford Cars	FGVC Aircraft	Flowers102	Food101	Oxford Pets	SUN397
CLIP [40, 21]	68.30	87.40	25.86	71.65	86.58	90.21	67.35
Retrieval + Finetuning	70.28	87.95	26.22	72.15	86.63	90.35	68.01
VLM [33]	69.35	87.88	28.54	72.11	86.31	90.24	67.73
CuPL [39]	70.25	88.63	29.64	72.32	86.20	91.16	70.80
Priming (Ours)	70.75	89.30	33.03	79.81	86.66	91.87	71.21
Priming + CuPL (Ours)	71.38	90.23	36.00	80.04	86.86	91.85	72.35

Related Directions

Related Directions

VISUAL CLASSIFICATION VIA DESCRIPTION FROM LARGE LANGUAGE MODELS

Sachit Menon, Carl Vondrick
Department of Computer Science
Columbia University

ABSTRACT

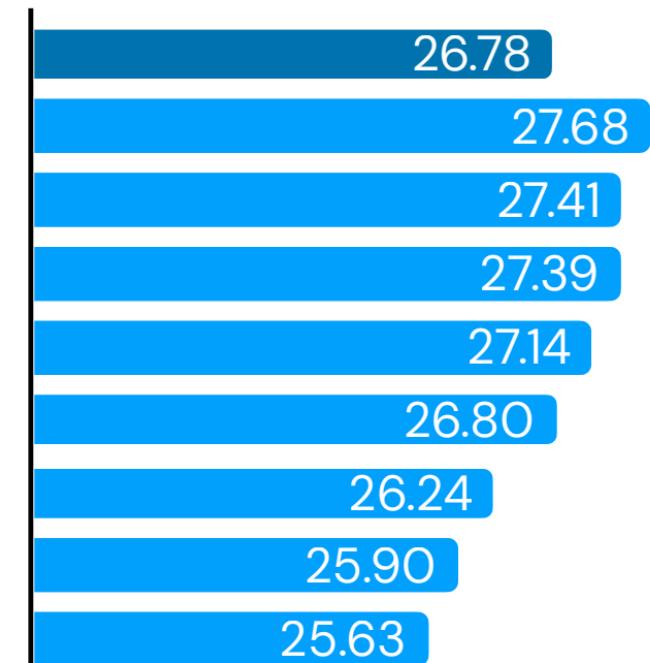
Vision-language models (VLMs) such as CLIP have shown promising performance on a variety of recognition tasks using the standard zero-shot classification procedure – computing similarity between the query image and the embedded words for each category. By only using the category name, they neglect to make use of the rich context of additional information that language affords. The procedure gives no intermediate understanding of why a category is chosen, and furthermore provides no mechanism for adjusting the criteria used towards this decision. We present an alternative framework for classification with VLMs, which we call classification by description. We ask VLMs to check for descriptive features rather than broad categories: to find a tiger, look for its stripes; its claws; and more. By basing decisions on these descriptors, we can provide additional cues that encourage using the features we want to be used. In the process, we can get a clear idea of what features the model uses to construct its decision; it gains some level of inherent explainability. We query large language models (e.g., GPT-3) for these descriptors to obtain them in a scalable way. Extensive experiments show our framework has numerous advantages past interpretability. We show improvements in accuracy on ImageNet across distribution shifts; demonstrate the ability to adapt VLMs to recognize concepts unseen during training; and illustrate how descriptors can be edited to effectively mitigate bias compared to the baseline.



Our top prediction: Hen
and we say that because...

Average

- two legs
- red, brown, or white feathers
- a small body
- a small head
- two wings
- a tail
- a beak
- a chicken

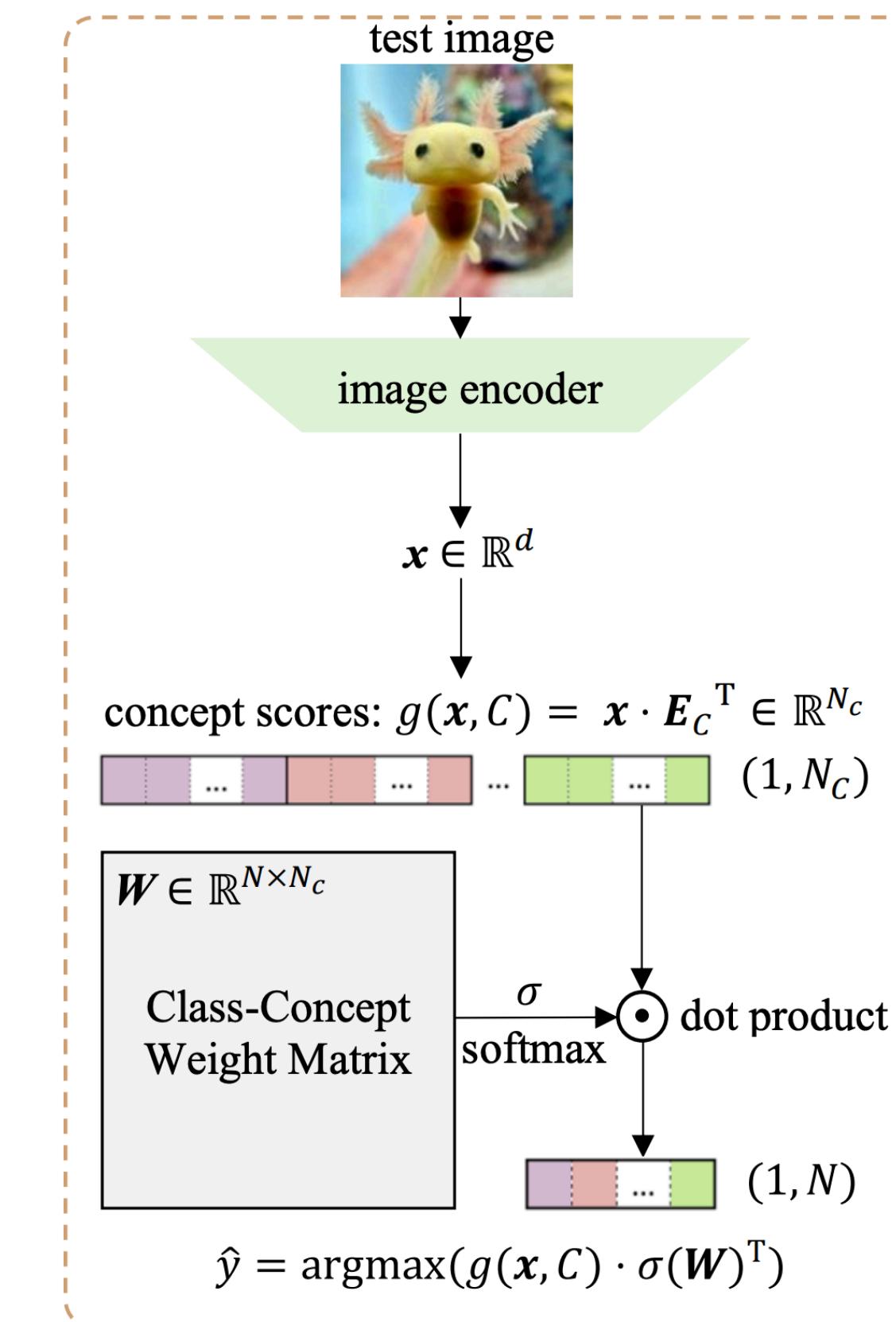


Related Directions

Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification

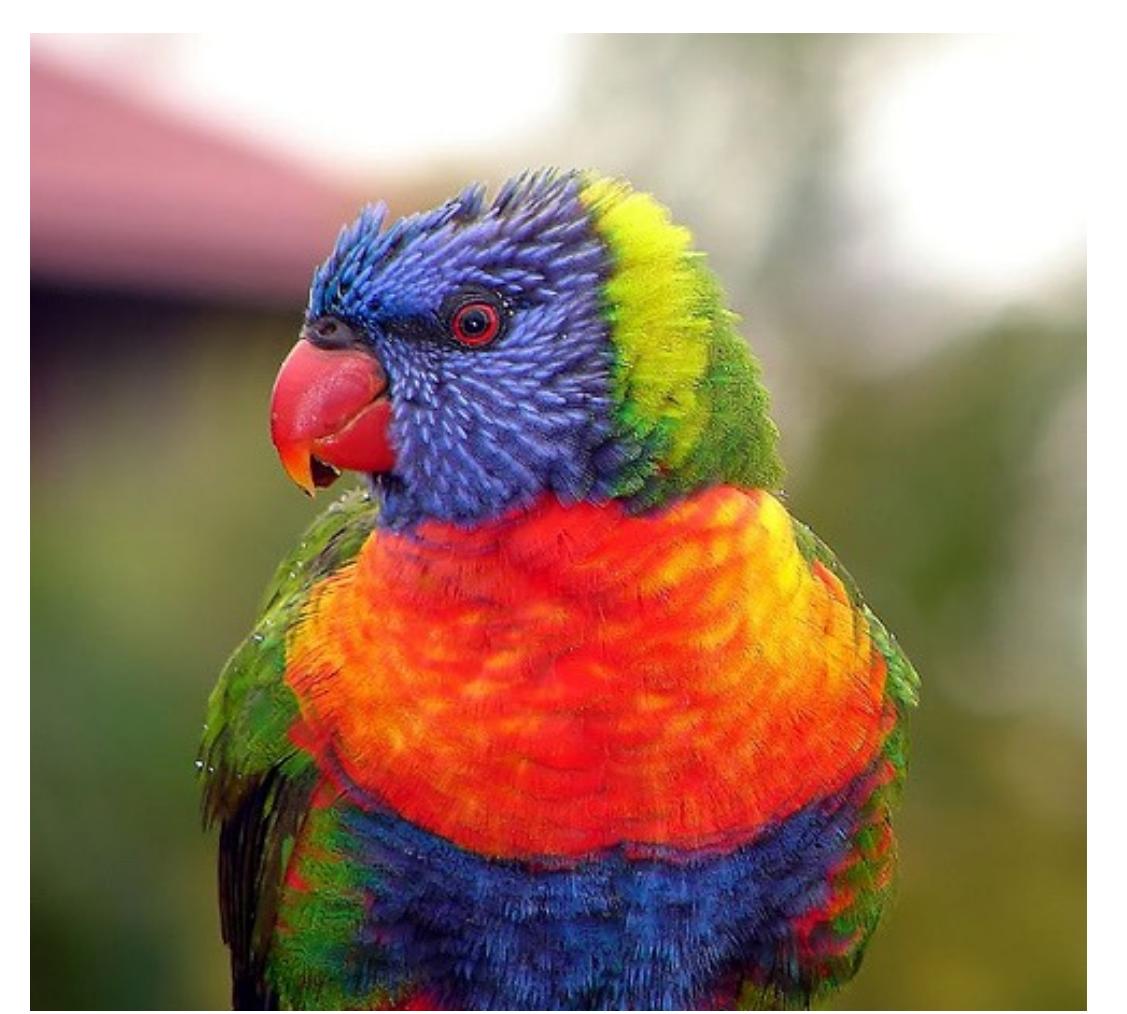
Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin,
Chris Callison-Burch, Mark Yatskar
University of Pennsylvania
`{yueyang1, artemisp, shzhou2, jindan, ccb, myatskar}@seas.upenn.edu`

Concept Bottleneck Models (CBM) are inherently interpretable models that factor model decisions into humanreadable concepts. They allow people to easily understand why a model is failing, a critical feature for high-stakes applications. CBMs require manually specified concepts and often under-perform their black box counterparts, preventing their broad adoption. We address these shortcomings and are first to show how to construct high-performance CBMs without manual specification of similar accuracy to black box models. Our approach, Language Guided Bottlenecks (LaBo), leverages a language model, GPT-3, to define a large space of possible bottlenecks. Given a problem domain, LaBo uses GPT-3 to produce factual sentences about categories to form candidate concepts. LaBo efficiently searches possible bottlenecks through a novel submodular utility that promotes the selection of discriminative and diverse information. Ultimately, GPT-3's sentential concepts can be aligned to images using CLIP, to form a bottleneck layer. Experiments demonstrate that LaBo is a highly effective prior for concepts important to visual recognition. In the evaluation with 11 diverse datasets, LaBo bottlenecks excel at few-shot classification: they are 11.7% more accurate than black box linear probes at 1 shot and comparable with more data. Overall, LaBo demonstrates that inherently interpretable models can be widely applied at similar, or better, performance than black box approaches.



Yang, Yue, et al. "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Final Thoughts



A lorikeet is a small to medium-sized parrot with a brightly colored plumage.

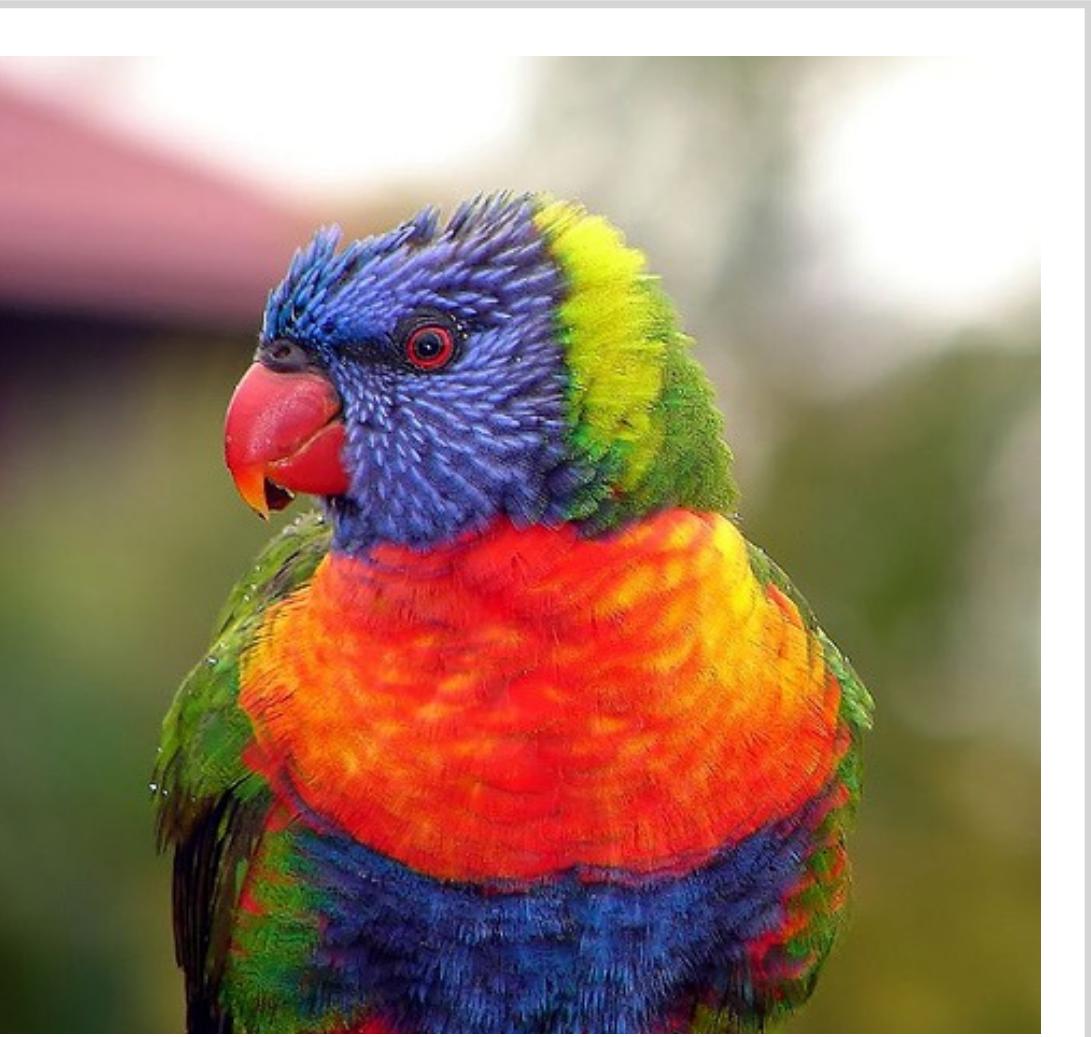
Final Thoughts

Neural Priming for Sample-Efficient Adaptation

Matthew Wallingford^{*†}, Vivek Ramanujan^{*†}, Alex Fang[†], Aditya Kusupati[†],
Roozbeh Mottaghi[†], Aniruddha Kembhavi^{†,o}, Ludwig Schmidt^{†,o}, Ali Farhadi[†]
[†]University of Washington, ^oAllen Institute for AI
{mcw244, ramanav}@cs.washington.edu

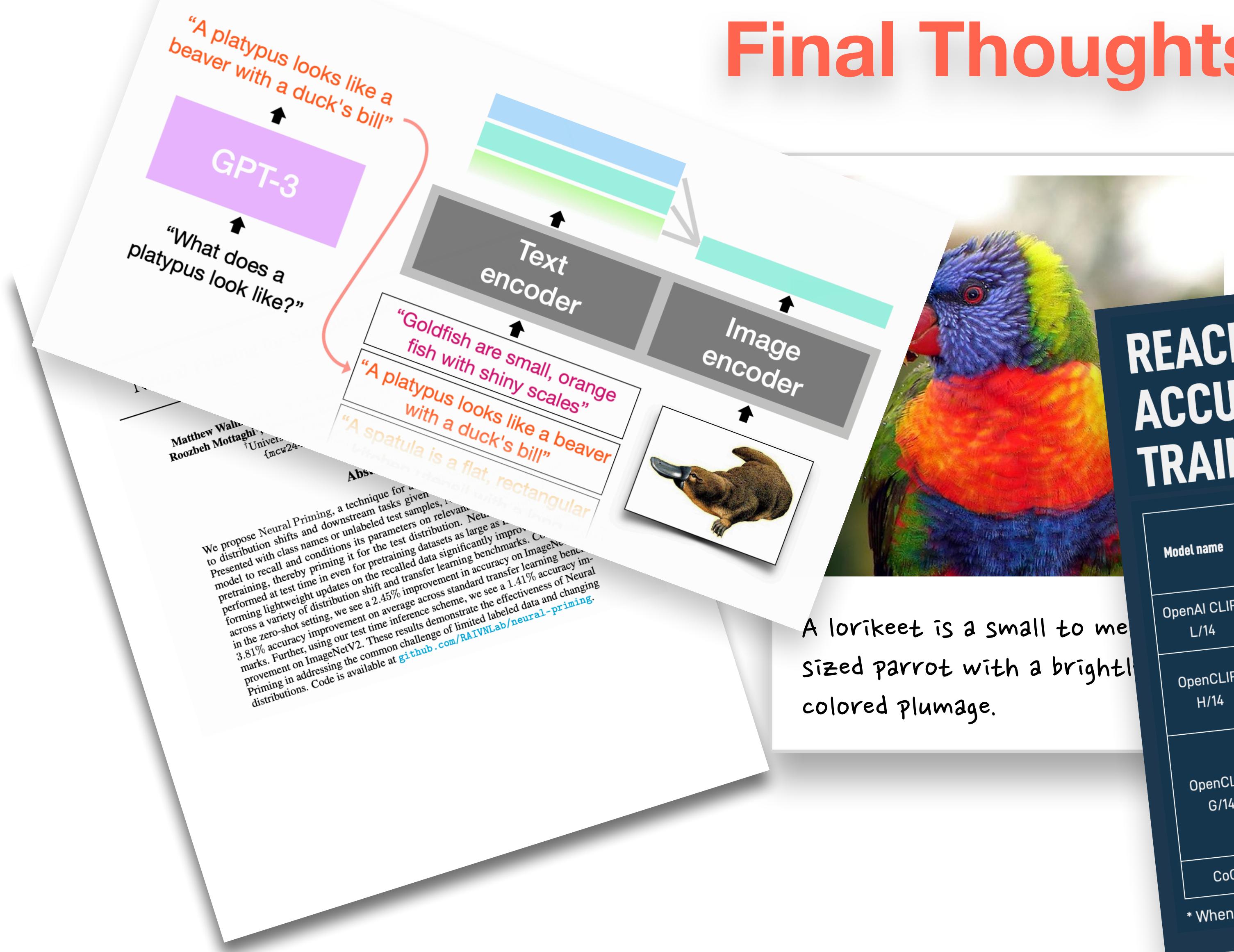
Abstract

We propose Neural Priming, a technique for adapting large pretrained models to distribution shifts and downstream tasks given few or no labeled examples. Presented with class names or unlabeled test samples, Neural Priming enables the model to recall and condition its parameters on relevant data seen throughout pretraining, thereby priming it for the test distribution. Neural Priming can be performed at test time in even for pretraining datasets as large as LAION-2B. Performing lightweight updates on the recalled data significantly improves accuracy across a variety of distribution shift and transfer learning benchmarks. Concretely, in the zero-shot setting, we see a 2.45% improvement in accuracy on ImageNet and 3.81% accuracy improvement on average across standard transfer learning benchmarks. Further, using our test time inference scheme, we see a 1.41% accuracy improvement on ImageNetV2. These results demonstrate the effectiveness of Neural Priming in addressing the common challenge of limited labeled data and changing distributions. Code is available at github.com/RAIVNLab/neural-priming.



A lorikeet is a small to medium-sized parrot with a brightly colored plumage.

Final Thoughts



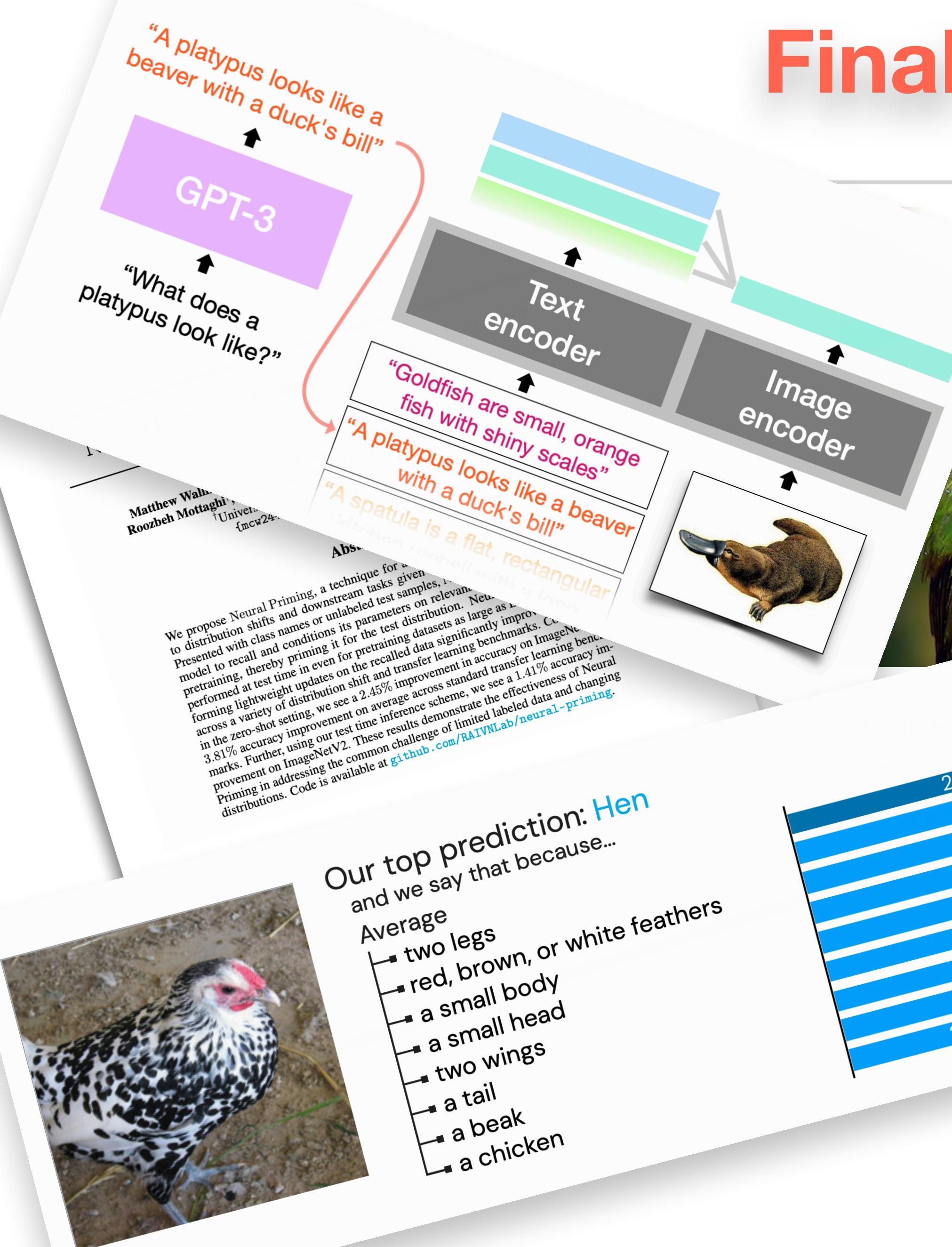
A lorikeet is a small to medium-sized parrot with a brightly colored plumage.

REACHING 80% ZERO-SHOT ACCURACY WITH OPENCLIP: VIT-G/14 TRAINED ON LAION-2B

Model name	Batch size	Samples seen	Text Params	Image params	ImageNet top1	Mscoco image retrieval at 5	Flickr30k image retrieval at 5
OpenAI CLIP L/14	32k	13B	123.65M	303.97M	75.4%	61.0%	87.0%
OpenCLIP H/14	79k	32B (16 epochs of laion2B)	354.0M	632.08M	78.0%	73.4%	94%
OpenCLIP G/14	160k	32B +unmasked fine-tune (details below)	694.7M	1844.9M	80.1%*	74.9%	94.9%
CoCa	66k	33B	1100M	1000M	86.3%**	74.2	95.7

* When using CuPL prompts instead of the standard prompts from OpenAI, the zero-shot accuracy is 80.3%.

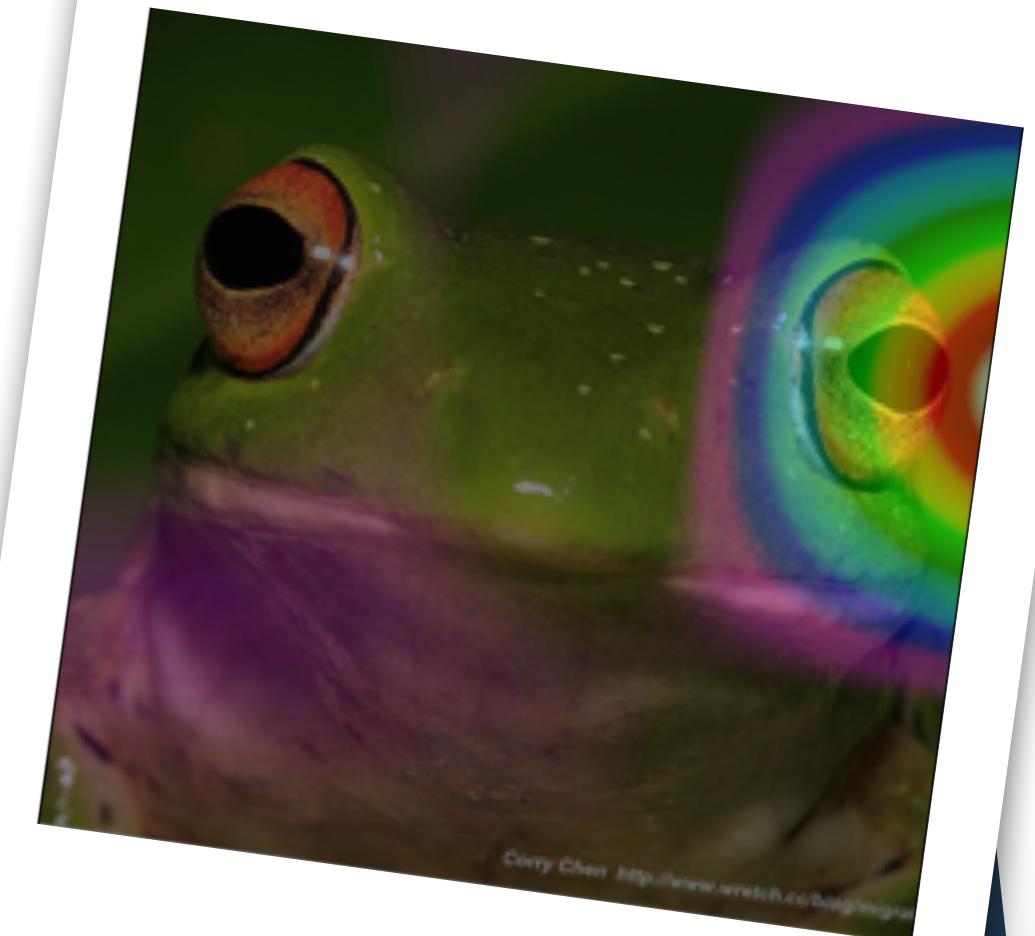
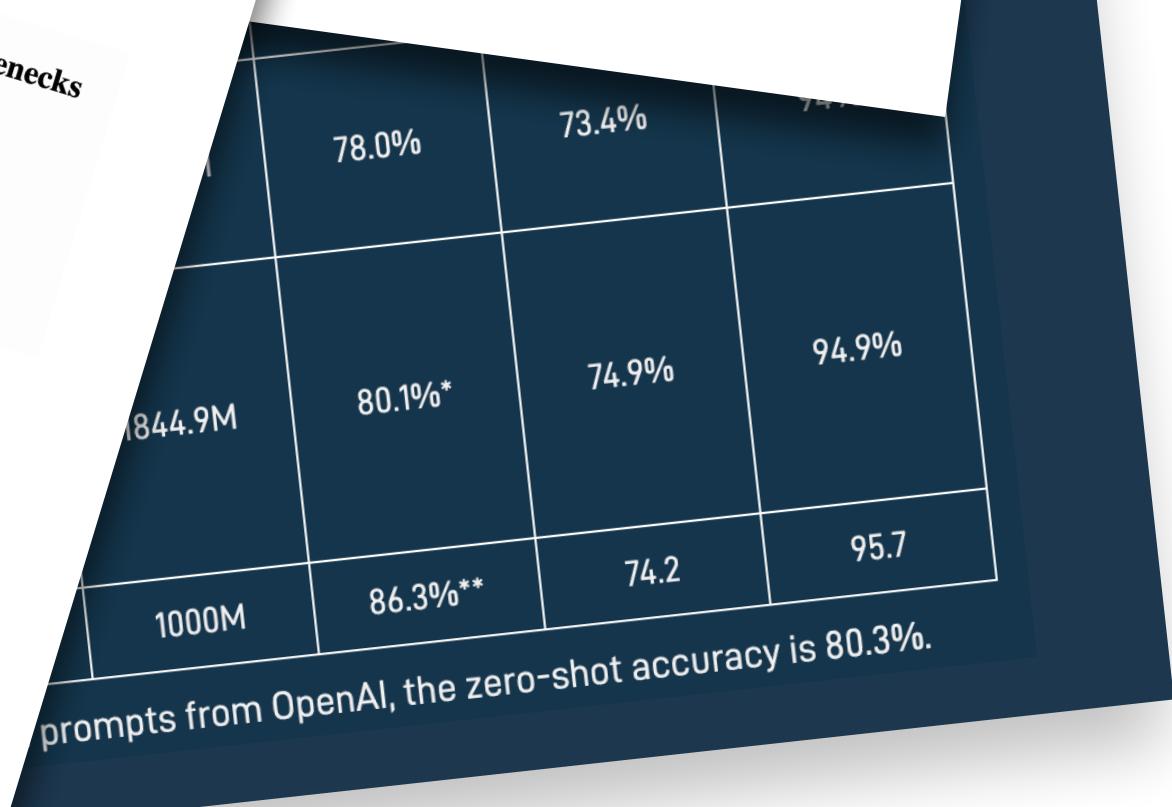
Final Thoughts



REACHING 80% ZERO-SHOT ACCURACY WITH OPENAI'S 100M

Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification
Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, Mark Yatskar
{yueyangl, artemisp, shzhou2, jindan, ccb, myatskar}@seas.upenn.edu

Concept Bottleneck Models (CBM) are inherently interpretable models that factor model decisions into human-readable concepts. They allow people to easily understand why a model is failing, a critical feature for high-stakes applications. CBMs require manually specified concepts and often under-perform their black box counterparts, preventing their broad adoption. We address these shortcomings and are first to show how to construct high-performance CBMs without manual specification of similar accuracy to black box models. Our approach, Language Guided Bottlenecks (LaBo), leverages a language model, GPT-3, to define a large space of possible bottlenecks. Given a problem domain, LaBo uses GPT-3 to produce factual



A tree frog looks like a small frog with es.

Thank You!



A photo of
Ian Covert



A photo of
Rosanne Liu



A photo of
Ali Farhadi

W UNIVERSITY *of* WASHINGTON

 DeepMind
 ML Collective

W UNIVERSITY *of* WASHINGTON